



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

工學博士學位論文

Next-generation sequencing based multi-omic
analysis of *Streptomyces* genome for deciphering
secondary metabolism

차세대 시퀀싱 기반 다중 오믹스 분석을 이용한
방선균 유전체의 분석 및 이차대사 생산의 이해

2015 년 8 월

서울대학교 大學院

化學生物工學部

金 鎮 玆

ABSTRACT

In this thesis, applications using next-generation sequencing (NGS) technology were employed to obtain genome-wide data, elucidating diverse cellular events of *Streptomyces* genome.

First, comparative genomic analysis using 17 completely sequenced genome of *Streptomyces* revealed that 2018 gene families constitute core genome of this genus, including 15 ortholog clusters of sigma factors, 22 ortholog clusters involved in cell division category and secondary metabolite genes related to stress protection.

Next, genome-wide binding of NdgR, a common transcriptional regulator involved in the biosynthesis of amino acids in *S. coelicolor*, was discovered by using chromatin immunoprecipitation coupled with high-throughput sequencing (ChIP-seq). The study showed that NdgR binds 19 genomic loci including upstream regions of most genes involved in branched-chain and sulfur-containing amino acids biosynthesis. For this experiment, tandem epitope tagging systems for *Streptomyces* genome engineering was developed, which can be applied to other transcription factors in *Streptomyces*. Further study revealed that NdgR maintains homeostasis of sulfur assimilation under thiol oxidative stress conditions.

In addition, genome architecture and dynamic expressions of mRNA and protein were uncovered by using multiple NGS tools, including TSS-seq, RNA-seq and ribosome profiling. Total 3926 transcription start sites were identified, indicating the length of 5' untranslated region of mRNA. This revealed that abundant existence of leaderless genes (~20%) and many of them were involved in transcription category. In particular, dynamic change of RNA and ribosome

protected mRNA fragment (RPF) level showed disparity between transcription and translation, indicating the existence of translational control. With the integration of multiple NGS data, the single-based resolution map of genome architecture and expression profiles of each secondary metabolite clusters were examined, which provides valuable information for manipulating secondary metabolite production. The enormous data generated in this thesis and methodologies can be applied to engineering of genetic circuits for the antibiotics synthesis in *S. coelicolor*.

Key words: *Streptomyces*, comparative genomics, transcriptomics, translomics, transcriptional regulation, next-generation sequencing, ChIP-seq, TSS-seq, RNA-seq, ribosome profiling

Student number: 2007-21181

CONTENTS

Abstract	i
LIST OF TABLES	vii
LIST OF FIGURES	viii

CHAPTER 1 Introduction

1.1 Genomic basis for secondary metabolite biosynthesis in <i>Streptomyces</i>	2
1.1.1 Genome sequencing of <i>Streptomyces</i>	2
1.1.2 Toward a systems level understanding of <i>Streptomyces</i> biology....	6
1.2 Next-generation sequencing technology	10
1.2.1 Emergence of next-generation sequencing.....	10
1.2.2 Next-generation sequencing methods.....	12
1.3 Applications of Next-generation sequencing technologies used in this thesis	19
1.3.1 Chromatin immunoprecipitation coupled with high-throughput sequencing (ChIP-seq).....	19
1.3.2 Strand-specific RNA sequencing (ssRNA-seq)	23
1.3.3 Differential RNA sequencing (dRNA-seq)	24
1.4 The scope of thesis	28

CHAPTER 2 Materials and methods

2.1 Bacterial strains and culture conditions	31
2.2 DNA manipulations	32
2.2.1 Construction of template plasmids for tandem myc tagging	32
2.2.2 Tandem epitope tagging to <i>Streptomyces coelicolor</i> transcription factors	32
2.3 Chromatin immunoprecipitation	33
2.4 RNA extraction	35
2.5 Directional RNA sequencing	35
2.6 Strand-specific RNA sequencing	38
2.7 NGS sequencing	40
2.8 Western blot analysis	40
2.9 Bioinformatic analysis	40
2.9.1 Pan-genome analysis	40
2.9.2 ChIP-seq data analysis	41
2.9.3 TSS identification and data analysis	41
2.9.4 RNA sequencing and ribosome profiling data processing	42

CHAPTER 3 Comparative genomics reveals the core and accessory genome of *Streptomyces* species

3.1 Pan-genome of 17 <i>Streptomyces</i>	46
3.2 Functional distribution of ortholog clusters	50

3.3 Core genome of 17 <i>Streptomyces</i> genome	53
3.4 Conclusion	61

CHAPTER 4 Genome-wide analysis of transcriptional regulatory network of NdgR in *Streptomyces coelicolor* using ChIP-seq

4.1 Construction of PCR-based tandem epitope tagging system for <i>Streptomyces</i> genome	63
4.2 Verification of tagging system using chromatin immunoprecipitation	65
4.3 Identification of in vivo NdgR binding regions by ChIP-seq	70
4.4 Sequence analysis of NdgR binding region	76
4.5 Functional classification of the NdgR regulon	77
4.6 Role of NdgR under thiol oxidative stress	81
4.7 Elucidation of NdgR regulatory logic	85
4.8 Conclusions	88

CHAPTER 5 Transcriptional and translational landscape of *Streptomyces coelicolor* genome

5.1 Integration of genome-wide data generated by Next-generation sequencing technology.....	91
5.2 High-resolution map of genetic organizational elements.....	98
5.3 Discrepancy in mRNA and protein expression.....	107
5.4 Genomic landscape of secondary metabolite genes in <i>S. coelicolor</i>	

.....	111
5.5 Conclusion	118
 CHAPTER 6 Conclusion & Further Suggestions	
Conclusion & Further Suggestions	119
 REFERENCES	122
 APPENDIX	
Appendix I The list of leaderless genes in <i>S. coelicolor</i>	142
Appendix II Transcriptional start sites in the secondary metabolite gene clusters of <i>S. coelicolor</i>	151
Appendix III RNA and RPF abundance of secondary metabolite genes in <i>S.</i> <i>coelicolor</i>	153
 ABSTRACT IN KOREAN	159

LIST OF TABLES

Table 1.1 <i>Streptomyces coelicolor</i> secondary metabolites.....	4
Table 1.2 Comparison of technical specifications of Next-generation sequencing platforms.....	18
Table 1.3 Applications of next-generation sequencing.....	20
Table 3.1 General genome features of 17 <i>Streptomyces</i> species used in this study.....	47
Table 3.2 COG distribution of ortholog clusters.....	51
Table 3.3 Conserved genes in 17 <i>Streptomyces</i> species.....	56
Table 3.4 Conserved genes involved in secondary metabolism in 17 <i>Streptomyces</i> species.....	58
Table 4.1 Genome-scale identification of NdgR binding regions.....	72
Table 4.2 The NdgR regulon genes.....	75
Table 5.1 Sequencing statistics.....	93

LIST OF FIGURES

Figure 1.1 Circular presentation of the <i>Streptomyces coelicolor</i> genome.....	3
Figure 1.2 Methodological omics tool kit.....	9
Figure 1.3 Workflow of conventional versus next-generation sequencing.....	11
Figure 1.4 The method used by the Roche/454 sequencer.....	13
Figure 1.5 Workflow of Illumina/Solexa sequencing.....	15
Figure 1.6 The ligase-mediated sequencing approach of the Applied Biosystems SOLiD sequencer.....	17
Figure 1.7 Schematic procedure of chromatin immunoprecipitation coupled with high-throughput sequencing.....	22
Figure 1.8 Schematic procedure of strand-specific RNA sequencing.....	25
Figure 1.9 Schematic procedure of differential RNA sequencing.....	26
Figure 2.1 Correlation between RNA-seq duplicate data.....	44
Figure 3.1 Pan-genome and core genome profiles.....	48
Figure 3.2 Venn diagram of specific genome in each species.....	49
Figure 3.3 Distribution of orthologous genes based on COG category.....	52
Figure 3.4 Location of core genome in each chromosome.	54
Figure 4.1 PCR-based tandem epitope tagging system for <i>S. coelicolor</i>	64
Figure 4.2 Confirmation of correct insertion of epitope tags into the chromosomal genes in <i>S. coelicolor</i>	66
Figure 4.3 Conservation of in vivo function of <i>scbR</i> -6× myc and <i>ndgR</i> -6× myc tagged strain.	68
Figure 4.4 Verification of tagging system using chromatin immunoprecipitaion	

.....	71
Figure 4.5 Genome-wide distributions of NdgR binding regions.....	73
Figure 4.6 Functional classification of genes in the NdgR regulon.....	78
Figure 4.7 Metabolic pathways directly regulated by NdgR.....	79
Figure 4.8 SigR-dependent transcription activation of the <i>ndgR</i> gene in <i>S. coelicolor</i>	83
Figure 4.9 The regulatory modes of NdgR.....	86
Figure 5.1 Integration of multiple high-throughput datasets mapped onto <i>S. coelicolor</i> genome.....	92
Figure 5.2 Comparison between previously known TSSs and the TSSs identified in this study.....	94
Figure 5.3 Growth phases of <i>S. coelicolor</i> and sampling point for this experiment.....	96
Figure 5.4 Differential expression of prodiginine gene cluster	97
Figure 5.5 Correlation between RNA-seq and Ribo-seq data.....	99
Figure 5.6 Novel sRNAs and ORFs within secondary metabolite gene clusters.	100
Figure 5.7 Identified TSSs classified by their positions.....	101
Figure 5.8 Distribution of 5' UTR lengths.....	102
Figure 5.9 Functional categories of leaderless genes.....	104
Figure 5.10 TE change distributions according to 5' UTR length and time points.....	105
Figure 5.11 Motif sequences drawn with total TSSs or the specific subset of total TSSs.....	106
Figure 5.12 Cluster analysis of abundance of mRNA and RPF by Pearson	

correlation.....	108
Figure 5.13 Expression pattern according to the median values of each cluster...	109
Figure 5.14 Functional enrichment analyses of the group of genes according to expression trend.....	110
Figure 5.15 TSSs in secondary metabolite gene clusters identified in this study.	112
Figure 5.16 Distribution of mRNA fold change and RPF fold change	114
Figure 5.17 Heatmap of whole secondary metabolite genes in <i>S. coelicolor</i>	115
Figure 5.18 LC-MS data of 5-hydroxyectoine.....	116

Chapter 1. Introduction

1.1 Genomic basis for secondary metabolite biosynthesis in *Streptomyces*

1.1.1 Genome sequencing of *Streptomyces*

Bioactive natural compounds have been widely used in medicinal and agricultural industry. Since the discovery of antituberculosis agents streptomycin from the culture broth of *Streptomyces griseus* by Waksman in 1944, the genus *Streptomyces* have been renowned for prolific source of such compounds, including antibacterials, immunosuppressants, antifungals, anticancer agents, insecticides and so on (Nett et al. 2009).

Streptomyces coelicolor A3(2) is the best studied model organism in the genus for the study of actinomycete chemistry and biology. It has linear chromosome with high G+C contents (72.1%), which was sequenced and annotated in 2002 (**Figure 1.1**) (Bentley et al. 2002), and contains 8,667,507 bp encoding 7769 genes and 56 pseudogenes. The linear SCP1 and circular SCP2 plasmids were sequenced later separately (Barrell et al. 2004; Brolle et al. 2003). One of the main reasons why *S. coelicolor* A3(2) was developed as a model genetic system is the production of pigmented antibiotics like blue-pigmented actinorhodin (ACT) and red-pigmented undecylprodigiosin (RED). Whole genome sequencing of this species revealed more than 20 secondary metabolite clusters and at present 30 gene clusters are confirmed or predicted for secondary metabolites biosynthesis (**Table 1.1**). These are various types of natural products such as polyketides, nonribosomal peptides, bacteriocins, terpenoids and so on.

Since the complete genome sequence of *S. coelicolor* was published, hundreds of genome sequences of *Streptomyces* have been publicly or privately opened

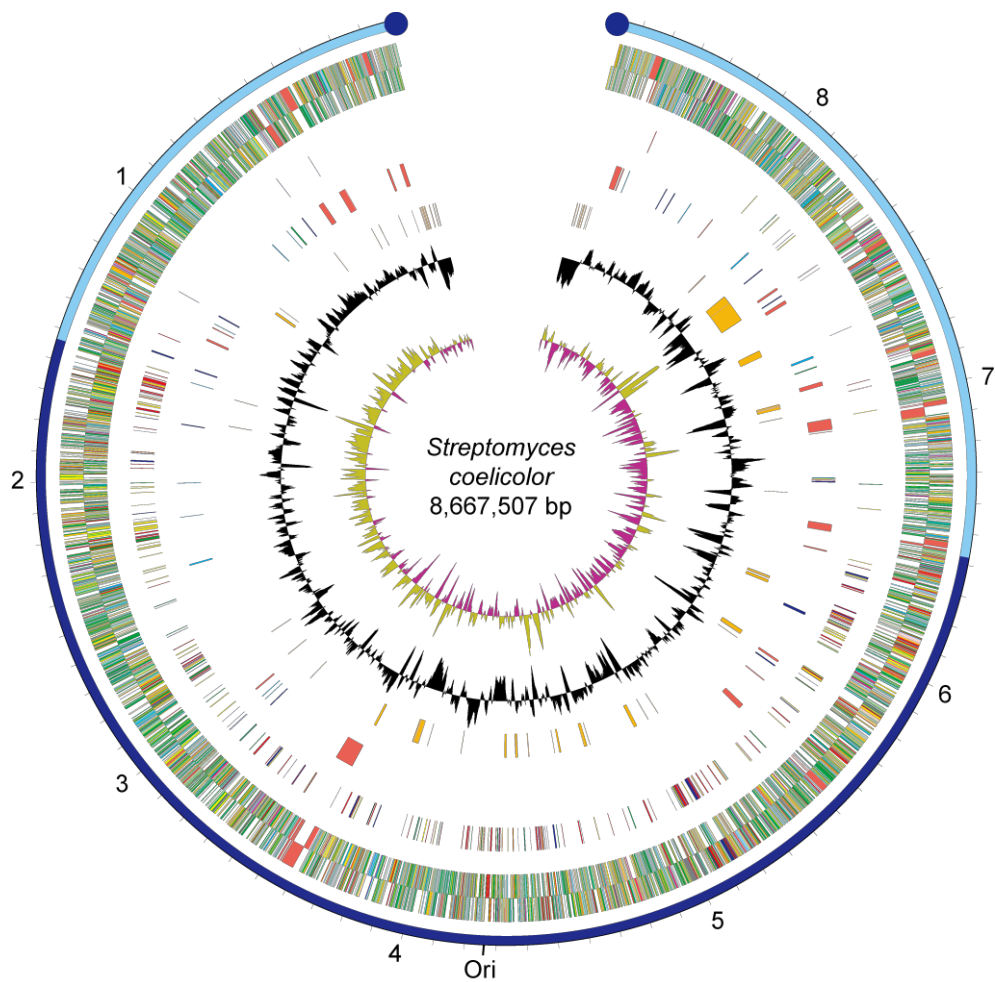


Figure 1.1 Circular presentation of the *Streptomyces coelicolor* genome (Bentley et al. 2002).

Table 1.1 *Streptomyces coelicolor* secondary metabolites (Craney et al. 2013).

<i>Secondary metabolite</i>	<i>Location</i>	<i>Type</i>
<i>Identified structures</i>		
Isorenieratene	SCO0185-0191	Terpenoid
Coelichelin	SCO0489-0499	NRP
THN/flaviolin	SCO1206-1208	PK—type III
5-Hydroxyectoine	SCO1864-1867	Cyclic amino acid
Desferrioxamine	SCO2782-2785	Tris-hydroxymate
CDA	SCO3210-3249	NRP
Actinorhodin	SCO5071-5092	PK—type II
Albaflavenone	SCO5222-5223	Terpenoid
Prodiginine	SCO5877-5898	Tripyrrole
Geosmin	SCO6073	Terpenoid
SCB1	SCO6266	g-Butyrolactone
Coelimycin P1	SCO6273-6288	PK—type I
Hopene	SCO6759-6771	PK—type III
Germicidin	SCO7221	PK—type III
2-Methylisoborneol	SCO7700-7701	Terpenoid
Methylenomycin	SCP1.228c-246	Cyclopentanoid
Methylfurans	SCP1.228c-246	Methylfurans
<i>Developmental secondary metabolites</i>		
SapB	SCO6681-6685	Lantibiotic
<i>Predicted structures (untested)</i>		
Eicosapentaenoic acid	SCO0124-0129	Fatty acid
Melanin	SCO2700-2701	Melanin
Bacteriocin	SCO0753-0756	Bacteriocin
Coelibactin	SCO7681-7691	NRP
<i>Unable to predict structures</i>		
Lantibiotic	SCO0267-0270	Lantibiotic
Lantibiotic	SCO6927-6932	Lantibiotic
PKS	SCO1265-1273	PK—type II
PKS	SCO6826-6827	PK—type II
PKS	SCO7669-7671	PK—type III
Siderophore	SCO5799-5801	—
Dipeptide	SCO6429-6438	—
Gray spore pigment	SCO5314-5320	PK—type II

including 17 completed genomes and many draft genomes as of October 2014 (Baranasic et al. 2013; Bentley et al. 2002; Hang et al. 2012; Nestor et al. 2014; Ohnishi et al. 2008; Pullan et al. 2011). From these efforts, the genomic backgrounds of numerous metabolic products were revealed. For example, the genome of *Streptomyces avermitilis*, a major producer of avermectin, harbors total of 37 secondary metabolite clusters including 11 polyketide synthase (PKS) gene clusters, eight nonribosomal peptides, six terpenoides and so on (Nett et al. 2009). Another pharmaceutically important species, *Streptomyces griseus*, encodes at least 36 gene clusters associated with the biosynthesis of secondary metabolites (Nett et al. 2009). Since the development of next-generation sequencing technology in mid-2000, the data size of genomic information has been rapidly increased. To date (September 2014), over 28000 species of bacterial genome sequence including 448 *Streptomyces* strains have been reported in NCBI genome database (<http://www.ncbi.nlm.nih.gov/genome/browse>). As the overflow of information, comparative genomics analyses between multiple genomes of individual species have been used to reveal extensive genomic intra-species diversity. Among the comparative analysis methods, pan-genome analysis is known for describing a bacterial species through characterizing core genome containing genes present in all strains (Tettelin and Massignani... 2005; Kettler et al. 2007; Rasko et al. 2008; Sugawara et al. 2013; Bottacini et al. 2014). This analysis results in the determination of core genome that probably encodes functions related to the basic biology and phenotypes of the species.

There have been several reports of comparative genomic studies of *Streptomyces* species (Jayapal et al. 2007; Nai-Hua and Ralph 2007; Zhou et al. 2012). The

analysis revealed a catalog of genome components and evolutionary history of *Streptomyces*. However, they used only a few strains for the analysis, which provided limited knowledge compared to the number of strains already sequenced.

Clearly, microbial genome sequencing and comparative genomics has provided the basis of understanding about chemistry and biology of *Streptomyces*, and future sequencing efforts will continue to exploit a number of valuable natural products.

1.1.2 Toward a systems level understanding of *Streptomyces* biology

As introduced, genomic information of secondary metabolite biosynthesis has been discovered by sequencing of numerous *Streptomyces*. However, sequencing the genome and discovering novel gene clusters is just the beginning; the genome sequencing is not enough for understanding the relationship between an organism's genome and its phenotype. The genome is organized by complex structural and functional elements, including elements that act at the protein and RNA levels and control dynamic expression of genes according to the circumstances.

To understand the functions and interactions of *Streptomyces* genes, it is necessary to characterize and quantify the gene products, mRNAs, proteins, metabolites and their interactions at the genome-wide level. For this reason, various 'omics' technologies have been generated, which enable quantitative monitoring of cellular components in a high-throughput manner, including transcriptomics, proteomics, metabolomics, interactomics and so on (**Figure 1.2**). The *Streptomyces* researchers have been using such omics tools to elucidate the structures and dynamics of biological systems.

The release of genomic data made it possible to carry out transcriptome study in

Streptomyces. The first transcriptome research of *Streptomyces* was accomplished by DNA microarray, was focused on the global expression change according to the growth phase from primary to secondary metabolism (Huang et al. 2001). Since then, DNA microarray has become one of the most powerful tools to the study of global gene expression and regulatory networks in *Streptomyces*. Global effects of sigma factors that coordinate the transcription initiation respond to environmental and physiological conditions were investigated (Lee et al. 2005). Moreover, transcriptome profiles affected by various mutants have been investigated using microarray (Lee et al. 2005; Kang et al. 2007; Hesketh et al. 2007; Lian et al. 2008; Chen et al. 2009). The chromatin immunoprecipitation (ChIP) experiments coupled with microarray (ChIP-chip) have been applied to study genome-wide mapping of transcriptional regulators in *Streptomyces*. Since the first ChIP-chip experiment for HspR in *S. coelicolor* was performed (Chen et al. 2009), the genome-wide identification of bindings of several transcription factors have been introduced (Hengst et al. 2010; Pullan et al. 2011; Kim et al. 2012a; Pérez-Redondo et al. 2012). Complementary to transcriptome studies, proteomic approaches make it possible to elucidate the global gene expression at protein level. The development of high-throughput technologies for detecting proteins such as two-dimensional gel electrophoresis (2D-GE), liquid chromatography coupled with tandem mass spectrometry (LC/MS/MS) and matrix-assisted laser desorption ionization-time-of-flight (MALDI-TOF) mass spectrometry, have allowed system-level researches of protein expression in *Streptomyces*. For example, proteome associated with primary and secondary metabolism was examined, including post-transcriptional modification (PTM) (Hesketh et al. 2002). Besides, a number of extracellular

proteome was detected (Kim et al. 2005) and global proteome changes under various stress conditions were measured by 2D-GE and mass spectrometry (Novotna et al. 2003).

The omics approaches provide a better understanding of the structure, dynamics and control mechanism of *Streptomyces* genome. As the development of next-generation sequencing (NGS) technology, various applications can be used for such omics studies more efficiently, generating enormous data from single experiment. In this thesis, multiple omics approaches using NGS technology was employed to study the systems level understanding of *Streptomyces* biology.

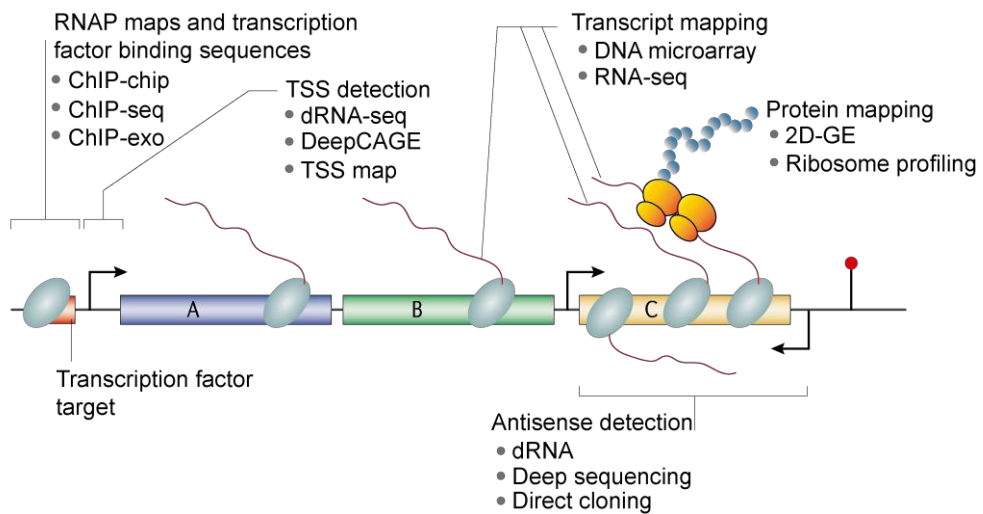


Figure 1.2 Methodological omics tool kit (Güell et al. 2011). Various methods that have been recently developed to measure different features of genome by mapping transcripts, antisense transcripts, transcription start sites (TSSs), proteins and protein-binding sites.

1.2 Next-generation sequencing technology

1.2.1 Emergence of next-generation sequencing

During the past few decades, Sanger sequencing method had dominated the sequencing technology (Sanger et al. 1977; Swerdlow et al. 1990; Hunkapiller et al. 1991). This led to a number of research accomplishments including complete human genome sequencing project (Consortium 2004). Despite of the big contributions in various fields, the limitations of automated Sanger sequencing were existed. Because the method was low throughput technology, enormous cost and time were required for obtaining huge data.

Since completion of the first human genome sequence, demand for cheaper and faster sequencing methods has increased greatly. This demand has driven the development of second generation sequencing methods. Finally, newer methods, referred to as next-generation sequencing (NGS) technologies, were introduced in the mid-2000 and realized in commercial products (e.g., 454 sequencing (Roche Applied Science), Solexa technology (Illumina), the SOLiD platform (Applied Biosystems) and so on) (Metzker 2009). These are commonly used the concept of cyclic-array sequencing which utilizes a dense array of DNA features by iterative cycles of enzymatic manipulation and imaging-based data collection (Mitra and Church 1999; Mitra et al. 2003). Although these platforms are quite diverse depending on the sequencing biochemistry, their workflows are conceptually similar (**Figure 1.3**). Library preparation is accomplished by random fragmentation of DNA, followed by *in vitro* ligation of common adaptor sequences. And clonally clustered amplicons are generated by several approaches, including emulsion PCR

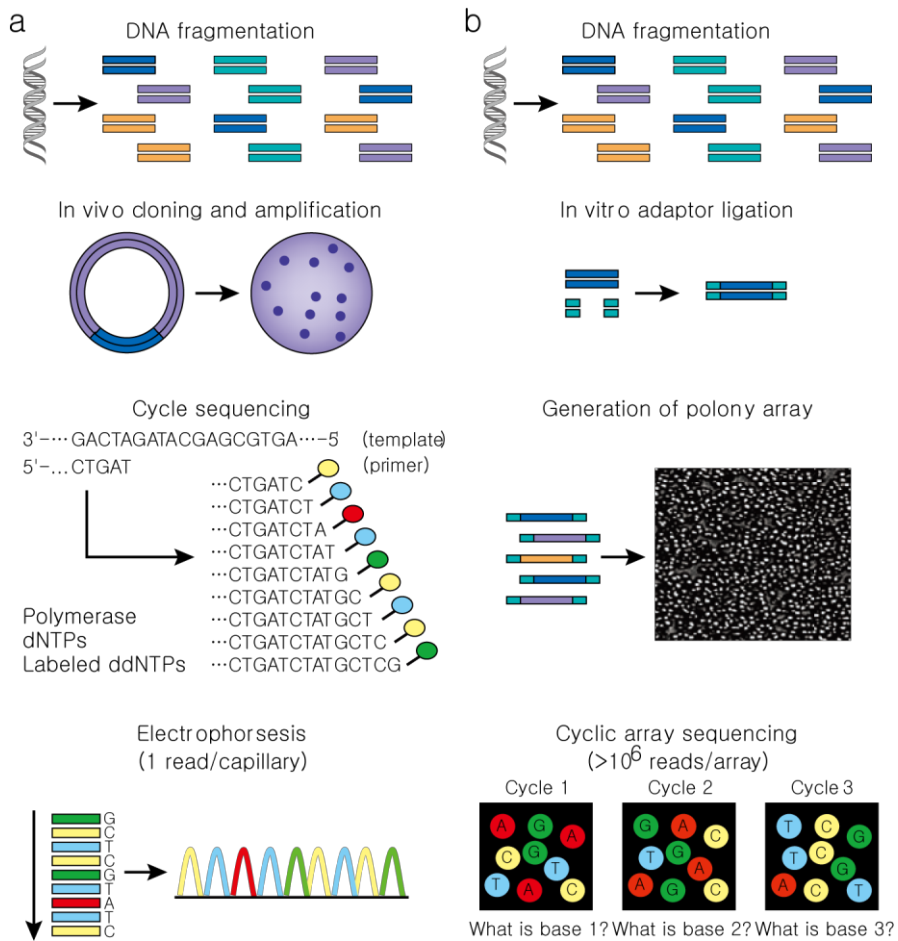


Figure 1.3 Workflow of conventional versus next-generation sequencing (Shendure and Ji 2008).

(Dressman et al. 2003) and bridge PCR (Adessi et al. 2000).

Global advantages of NGS, relative to Sanger sequencing, include the following: (1) *in vitro* construction of a sequencing library, followed by *in vitro* clonal amplification to generate sequencing features, enables parallel sequencing. (2) Array-based sequencing enables a much higher degree of parallelism. (3) Because array features are immobilized to a planar surface, they can be enzymatically manipulated by a single reagent volume.

There are some disadvantages of NGS relative to Sanger sequencing such as short read-length and raw accuracy. However, these limitations will continue to be overcome by technical improvement as conventional sequencing progressed gradually over three decades to reach its current level of technical performance.

1.2.2 Next-generation sequencing methods

Three platforms for massively parallel DNA sequencing read production are used widespread in the marketplace according to the technical differences: 454 sequencing (Roche), Solexa technology (Illumina), SOLiD platform (Applied Biosystems). These techniques constitute various processes including template preparation, sequencing, imaging and data analysis.

454 pyrosequencing. The 454 system was the first NGS platform available as a commercial product (Margulies et al. 2005). In this approach, libraries that are a mixture of short, adaptor-flanked fragment are amplified by emulsion PCR, with amplicons captured to the surface of 28- μ m beads (**Figure 1.4**). After breaking the

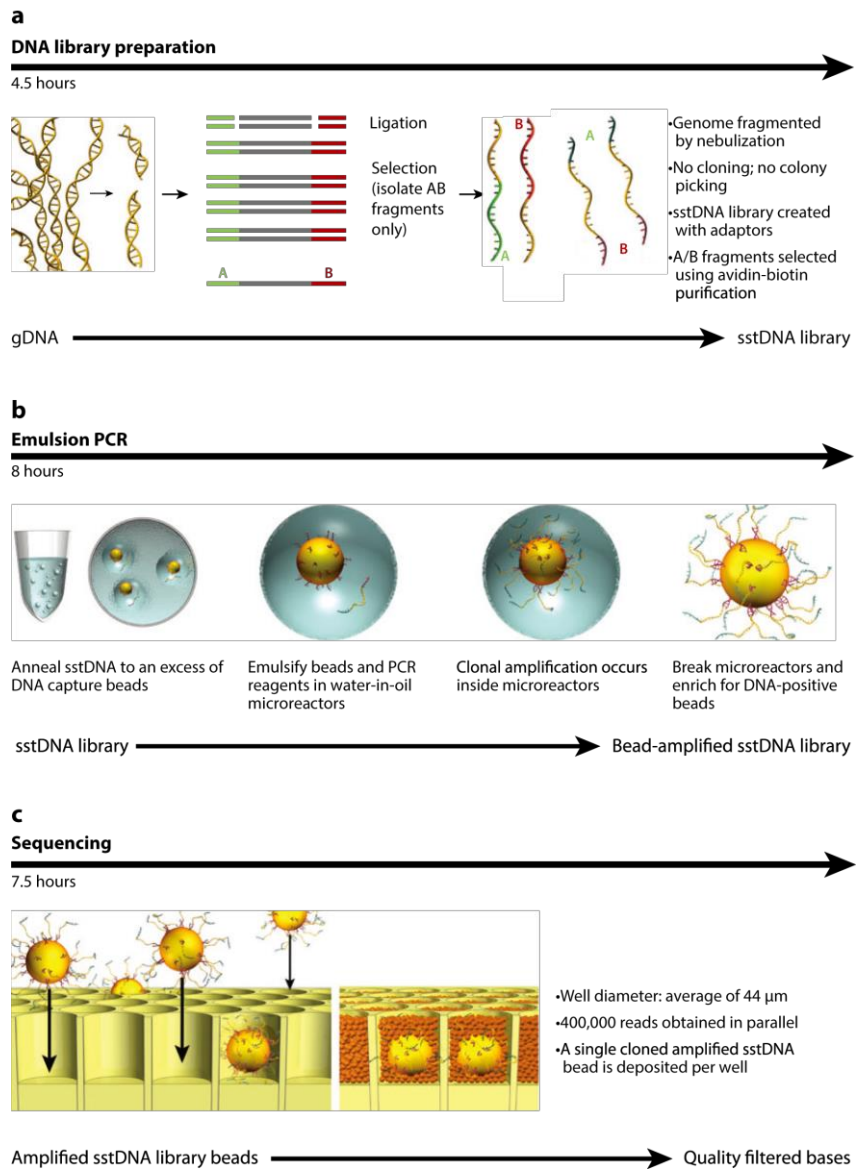


Figure 1.4 The method used by the Roche/454 sequencer (Mardis 2008).

emulsion, beads are treated with denaturant to remove untethered strands, and then subjected to a hybridization-based enrichment for amplicon-bearing beads. A sequencing primer is hybridized to the universal adaptor at the appropriate position and orientation, that is, immediately adjacent to the start of unknown sequence by pyrosequencing method (Ronaghi et al. 1996). The key advantage of this technology is long read-length. So, it is better to use where long read-lengths are critical such as de novo assembly and metagenomics. However, a major limitation is high error rate when homopolymers exist in the sequence.

Solexa technology. Libraries can be constructed as a mixture of adaptor-flanked fragments up to several hundred base-pairs in length. Amplified sequencing features are generated by bridge PCR (**Figure 1.5**). In this approach, both forward and reverse PCR primers are tethered to a solid substrate by a flexible linker. And amplicons are clustered to a single physical location on an array. After cluster generation, the amplicons are single stranded and a sequencing primer is hybridized to a universal sequence flanking the region of interest. And they are sequenced by using the four-color cyclic reversible termination method. The major advantage of Solexa technology is the amount of outputs. It can be used for the requirement of deep sequencing such as expression profiling. Read-length are relatively short (50~300 bp), however, this have been overcome by technical improvement.

AB SOLiD. This platform also uses sequencing features generated by emulsion PCR and sequencing by ligation. Clonally amplified 1- μ m beads are used to

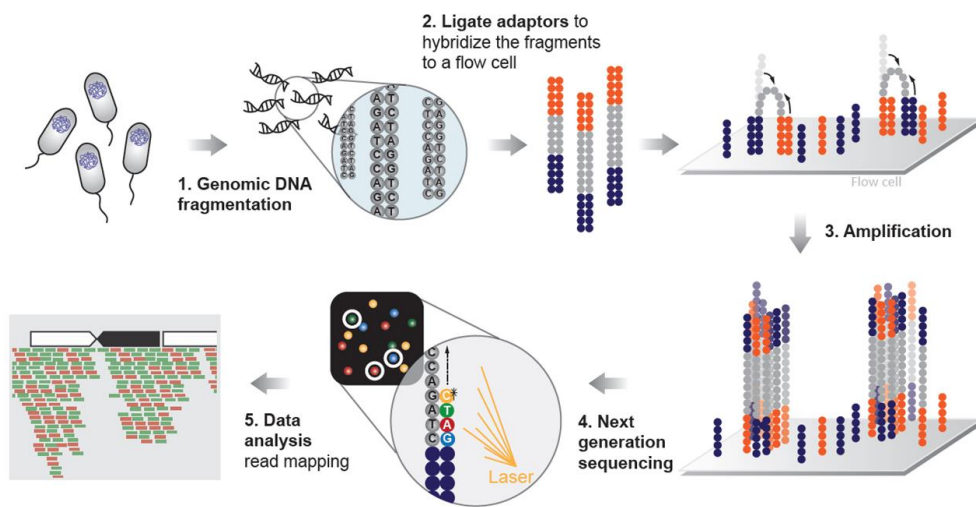


Figure 1.5 Workflow of Illumina/Solexa sequencing.

generate a disordered, dense array of sequencing features. Sequencing is performed with a ligase, rather than a polymerase. Each sequencing cycle introduces a partially degenerate population of fluorescently labeled octamers (**Figure 1.6**). The cost of the instrument is lower than other NGS instruments. However, the read-length may be significantly limiting.

Recent sequencing platforms. The sequencing technology have been evolving rapidly and three major new sequencing platforms were released in 2011: Ion Torrent Personal Genome Machine (PGM), Pacific Biosciences (PacBio) RS and Illumina MiSeq. PGM uses semiconductor technology detecting the protons released as nucleotides are incorporated during synthesis (Rothberg et al. 2011). PacBio have developed a process enabling single molecule real time (SMRT) sequencing (Eid et al. 2009). Each sequencing technology has clear advantages for particular applications over others (**Table 1.2**).

In recent years, the sequencing industry has been dominated by Illumina, Genome Analyzer and more recently HiSeq 2000 have set the standard for high throughput massively parallel sequencing. But, in 2011 Illumina released a lower throughput fast-turnaround instrument, the MiSeq, aimed at smaller laboratories and the clinical diagnostic market. The sequencing data resulted in this thesis was generated by using MiSeq platform due to its convenient handling in the laboratory.

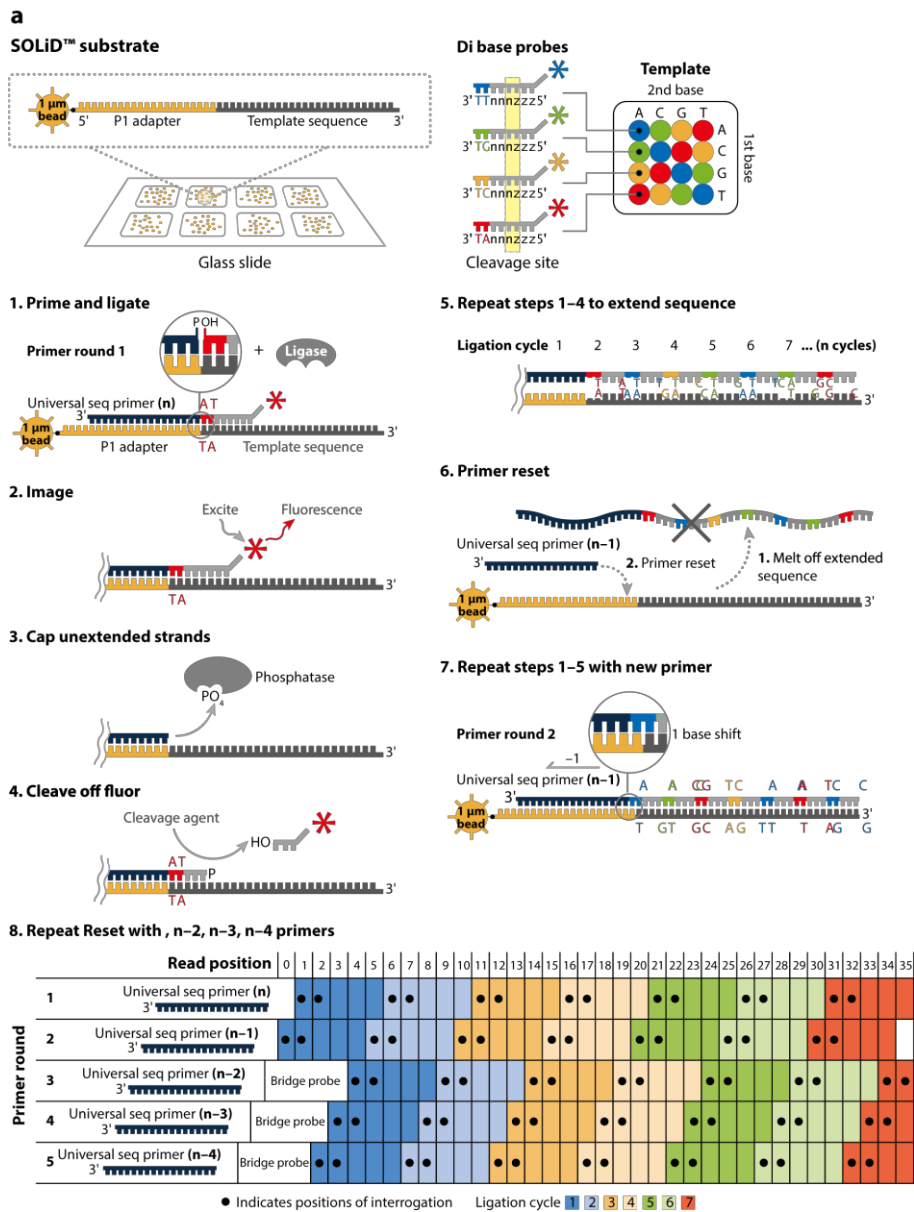


Figure 1.6 The ligase-mediated sequencing approach of the Applied Biosystems SOLiD sequencer.

Table 1.2 Comparison of technical specifications of Next-generation sequencing platforms (Quail et al. 2012).

Platform	Illumina Miseq	Ion Torrent PGM	PacBio RS	Illumina GAIIx	Illumina HiSeq 2000
Instrument Cost	\$128K	\$80K	\$695K	\$256K	\$654K
Sequence yield per run	1.5-2Gb	20-50Mb on 314 chip, 100-200Mb on 316 chip, 1Gb on 318 chip	100Mb	30Gb	600Gb
Run time	27 hours	2 hours	2 hours	10 days	11 days
Reported Accuracy	Mostly>Q30	Mostly Q20	<Q10	Mostly>Q30	Mostly>Q30
Observed Raw Error Rate	0.80%	1.71%	12.86%	0.76%	0.26%
Read length	up to 150 bases	~200 bases	Average 1500 bases	up to 150 bases	up to 150 bases
Paired reads	Yes	Yes	No	Yes	Yes
Insert size	up to 700 bases	up to 250 bases	up to 10kb	up to 700 bases	up to 700 bases
Typical DNA requirements	50-1000 ng	100-1000 ng	~1µg	50-1000 ng	50-1000 ng

1.3 Applications of Next-generation sequencing technologies used in this thesis

This revolutionized sequencing technology has changed one's imagination about the limitation of scientific approaches (**Table 1.3**). For example, whole genome sequences of various organisms have been opened with rapid and lower cost compared to the era of first generation technology. This has allowed large-scale comparative genomics and evolutionary studies. In addition, genome-wide measurement of gene expression has been possible even in the single-based resolution, which can detect novel transcripts without prior knowledge of a particular gene. In this thesis, these tools have been applied to elucidating the genomic background of *Streptomyces* biology. Each of the NGS applications used in this study will be introduced in the following sections.

1.3.1 Chromatin immunoprecipitation coupled with high-throughput sequencing (ChIP-seq)

Identification of DNA-protein interaction is essential for understanding transcriptional regulation (Balleza and López-Bojorquez 2009). Especially, genome-wide mapping of DNA-protein interaction is vital for deciphering the gene regulatory network under certain biological conditions. The main tool for investigating these interaction mechanisms *in vivo* is chromatin immunoprecipitation (ChIP) (Orlando 2000). In a ChIP technique, DNA-binding protein is crosslinked to DNA *in vivo* by treating cells with formaldehyde. And chromatin is fragmented by sonication, which are generally in the 200-600 bp range. Then, antibodies are used to extract specific proteins that are interested in

Table 1.3 Applications of next-generation sequencing (Shendure and Ji 2008).

<i>Category</i>	<i>Examples of applicaitons</i>
Complete genome resequencing	Comprehensive polymorphism and mutation discovery in individual human genomes
Reduced representation sequencing	Large-scale polymorphism discovery
Targeted genomic resequencing	Targeted polymorphism and mutation discovery
Paired end sequencing	Discovery of inherited and acquired structural variation
Metagenomic sequencing	Discovery of infectious and commensal flora
Transcriptome sequencing	Quantification of gene expression and alternative splicing; transcript annotation; discovery of transcribed SNPs or somatic mutations
Small RNA sequencing	microRNA profiling
Sequencing of bisulfite-treated DNA	Determining patterns of cytosine methylation in genomic DNA
Chromatin immunoprecipitation–sequencing (ChIP-Seq)	Genome-wide mapping of protein-DNA interactions
Nuclease fragmentation and sequencing	Nucleosome positioning
Molecular barcoding	Multiplex sequencing of samples from multiple individuals
Ribosome profiling	Annotation of translated sequences

the goal of research. DNA fragments that are bound to the proteins are enriched for characterization.

Development of high-density tiling microarray allowed the enriched DNA fragments obtained from ChIP to be identified by hybridization (ChIP-chip), therefore enabling a genome-wide view of DNA-protein interactions (Ren et al. 2000). However, microarray only detects the level over the signal-to-noise ratio and low resolution is the limitation of this technique.

The emergence of next-generation sequencing technology has overcome the drawbacks of hybridization technique. The enriched DNA fragments from ChIP experiments are massively sequenced in a single run, allowed large experiments that could only be imagined. This is called chromatin immunoprecipitation coupled with high-throughput sequencing (ChIP-seq) (**Figure 1.7**) (Robertson et al. 2007). ChIP-seq offers many advantages over ChIP-chip, which has higher resolution, greater coverage and larger dynamic range than ChIP-chip (Park 2009). Therefore, the precise mapping of protein-binding region is allowed and more accurate binding motif can be identified.

Recently, it has been used to elucidate the genome-wide binding of regulators in *Streptomyces* (Higo et al. 2012; Bush et al. 2013; Al-Bassam et al. 2014). In this thesis, ChIP-seq technique was employed to elucidate the regulatory network of NdgR in *S. coelicolor*. NdgR is highly conserved among *Streptomyces* species as well as other actinomycetes such as *Mycobacterium* and *Corynebacterium* (Yang et al. 2009b; Kim et al. 2012b). NdgR and its orthologs are located adjacent to *leuCD*, which encodes isopropylmalate dehydratase, and have been identified as its transcriptional regulator. In addition, the last step of methionine biosynthesis has

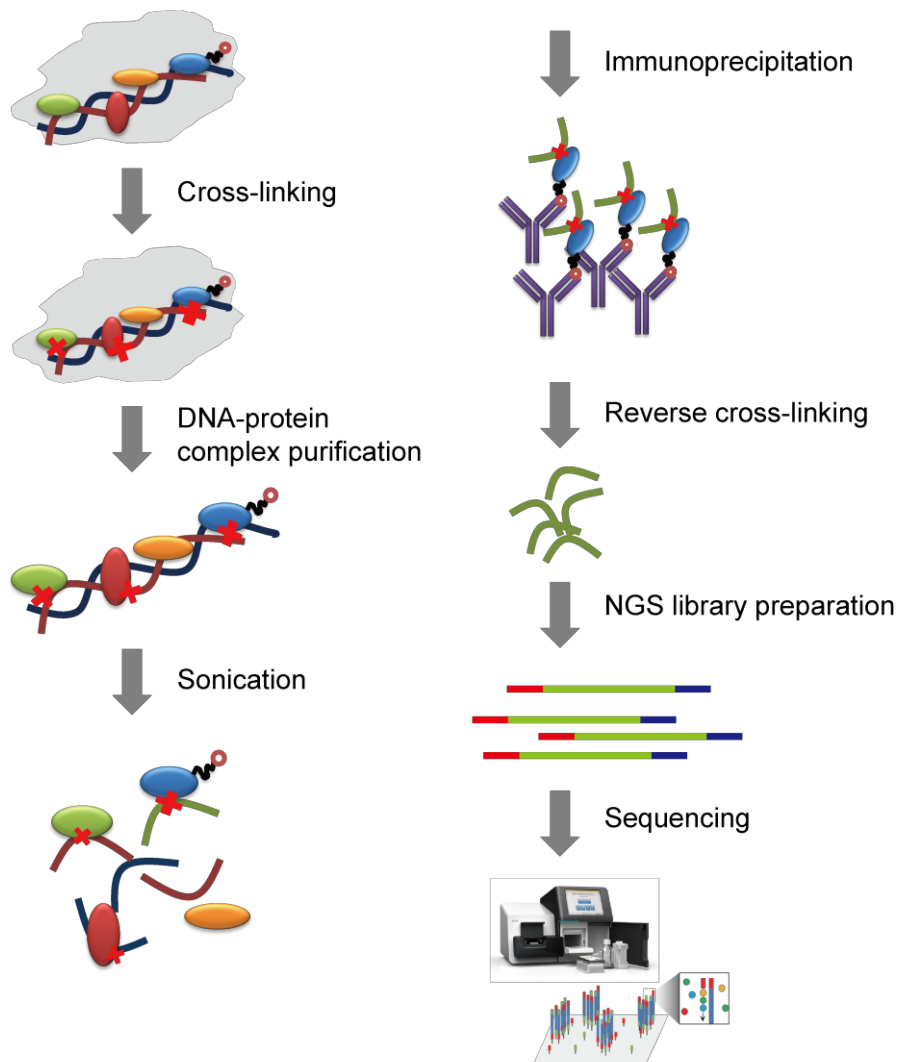


Figure 1.7 Schematic procedure of chromatin immunoprecipitation coupled with high-throughput sequencing (ChIP-seq).

been identified as a regulatory target of NdgR (Kim et al. 2012b). Regulation of *ndgR* orthologs in other bacteria has also been revealed. For example, AreB in *Streptomyces clavuligerus* controls the biosynthesis of leucine and secondary metabolites (Santamarta et al. 2007). In *Corynebacterium glutamicum*, LtbR has been characterized as a regulator involved in leucine and tryptophan biosynthesis (Brune et al. 2007).

For the identification of direct binding targets of DNA-binding proteins of *Streptomyces*, we developed a versatile PCR-based tandem epitope tagging tool (Chapter 4). Using this tool, we constructed a *S. coelicolor* harboring a 6× myc-tagged NdgR that was successfully utilized for the immunoprecipitation of NdgR-DNA complexes. This experiment revealed that NdgR regulates not only *leuCD* but also most of genes involved in leucine biosynthesis in *S. coelicolor*. Despite the previous studies, the DNA-binding locations of NdgR under physiologically relevant conditions have not been described at a high resolution, and on a genome-wide scale. In chapter 4, we investigated the NdgR regulon in vivo using chromatin immunoprecipitation coupled with high-throughput sequencing (ChIP-seq) and further explored the responses of the NdgR regulatory network to thiol oxidative stress. The regulatory roles of NdgR in *S. coelicolor* will be discussed.

1.3.2 Strand-specific RNA sequencing (ssRNA-seq)

A revolution in transcriptomics study started when the DNA microarray was developed for quantifying global RNA expression (Schena et al. 1995). Several years later, microarrays containing high-density tiled probes (termed tiling arrays) that cover the entire genome of an organism were developed and the first

comprehensive transcriptome map for *Escherichia coli* and several other bacteria were introduced using such tiling arrays (Selinger et al. 2000). However, the tiling array detected only the transcripts in the levels that are above background noise and cannot provided single-based resolution of transcription. Thus, this system has a limitation to detect exact quantity of transcription. In 2008, RNA-seq was introduced following the development of deep sequencing technique (**Figure 1.8**) (Lister et al. 2008). It sequences cDNA generated from RNA preparations. This technology has overcome the limitations of tiling arrays: it reduces signal-to-noise ratio and provides higher dynamic range. And it sequences transcript as a single-based resolution (Wang et al. 2009).

This revolutionary method has revealed various functional genomic elements and their regulatory roles in bacteria (Rotem Sorek 2009). First, gene structure can be reannotated by transcriptome analysis. Most sequenced genomes are annotated by gene-prediction software, but they are error-prone and fail to detect noncoding RNAs. New genes and antisense transcriptions can also be discovered without previous information. Second, transcriptional expression can be measured in genome-scale and single-based resolution. This makes it possible to detect RNA-based regulation in bacteria.

1.3.3 Differential RNA sequencing (dRNA-seq)

RNA sequencing techniques have been modified by the goal of researches. One of them is the detection of transcription start sites (TSSs) (**Figure 1.9**). Transcription start sites can be detected by selecting only primary transcripts that have a 5'-triphosphate (Sharma et al. 2010). TSS profiling revealed complex transcription

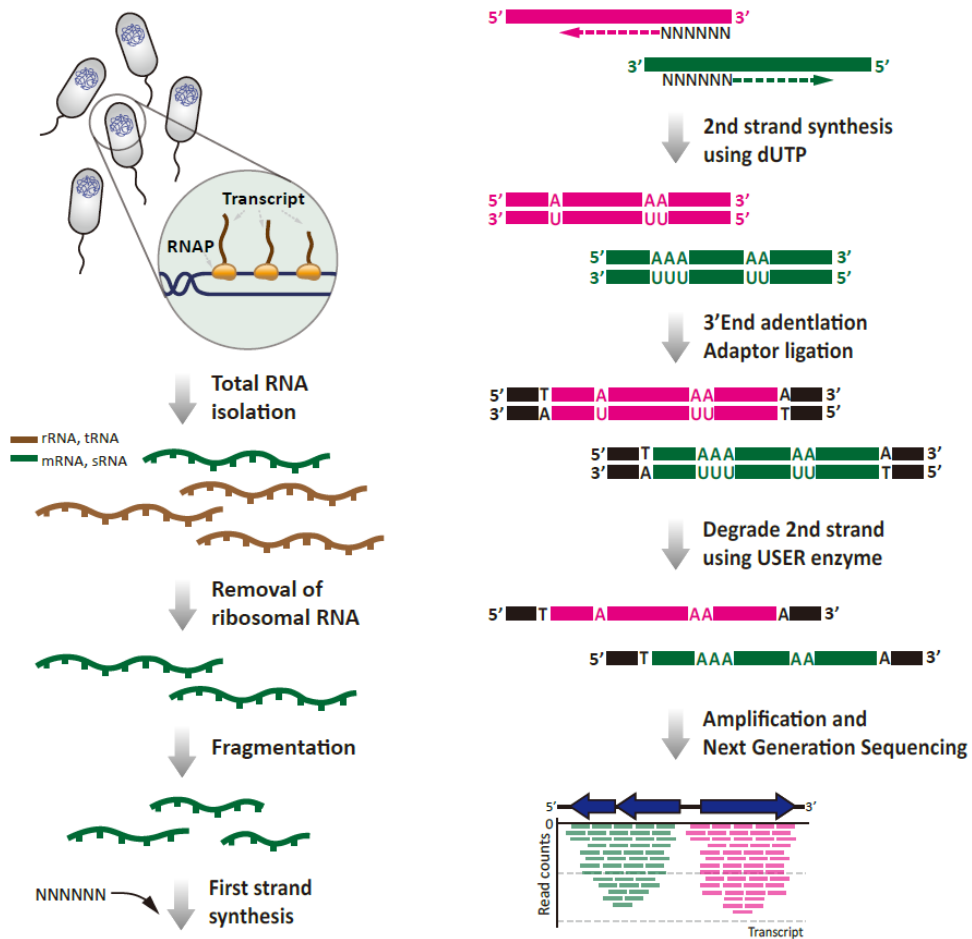


Figure 1.8 Schematic procedure of strand-specific RNA sequencing (ssRNA-seq).

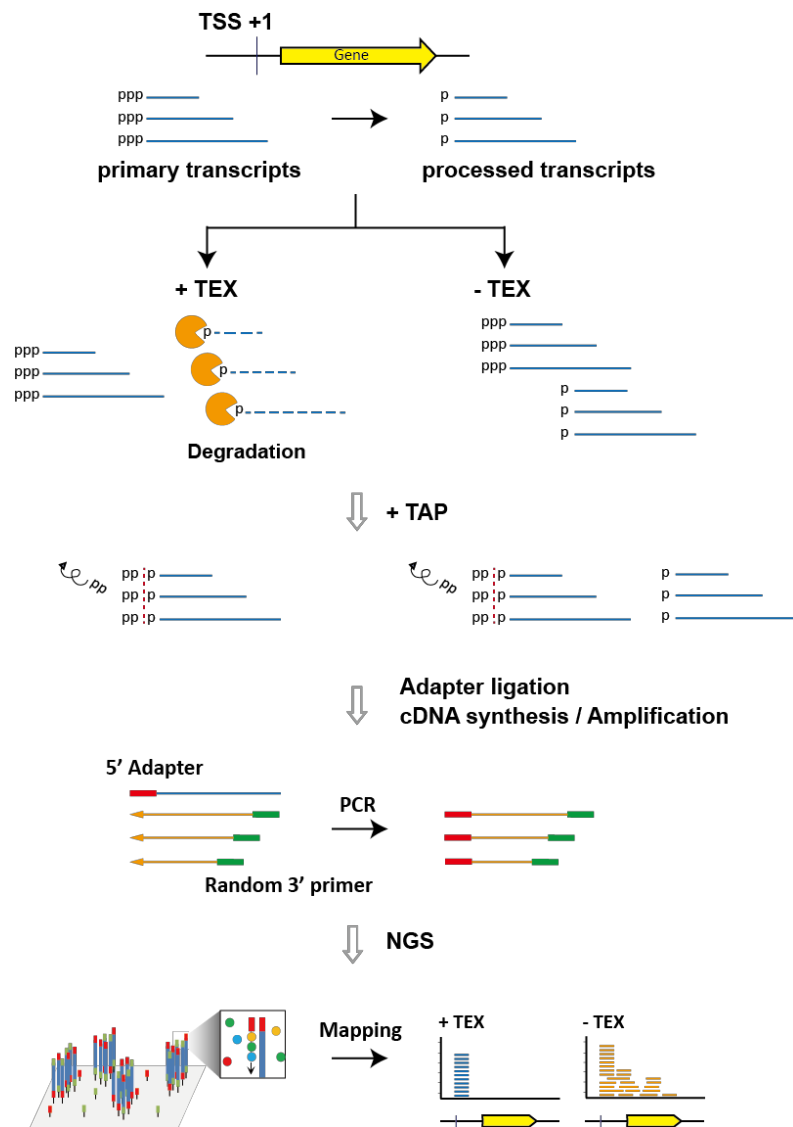


Figure 1.9 Schematic procedure of differential RNA sequencing (dRNA-seq or TSS-seq).

initiation of bacterial gene expression. For example, multiple TSSs of single gene were detected. It means that the promoter regions of bacterial genes were bound by multiple sigma factors that recruit RNA polymerases under the growth phases or specific environments. Also, internal TSSs within operons were detected, thereby adding plasticity to the operon. This revealed that consecutive genes within operons do not have the same expression level. And the expression can be changed by physiological situation similar to eukaryotes. Furthermore, identification of TSSs determined the sequence and length of untranslated regions (UTRs). Some genes have riboswitches that regulates translation of corresponding genes by interacting with small molecules or other biological compounds. And the existences of leaderless mRNAs that have no UTRs were founded frequently.

This information provides the genomic architecture that represents the region of intact transcription of RNA in bacteria. Using this tool, precise mapping of *Streptomyces* can be established.

1.4 The scope of thesis

The purpose of this thesis is uncovering functional element of *Streptomyces* genome beyond the genome sequence itself. To attain this end, I focus on the genome-wide analyses of the dynamic aspects of *Streptomyces* genes such as transcription, translation and DNA-protein interactions, not only the static aspects of the genomic information such as DNA sequences or structures.

In chapter 3, the genomic properties of *Streptomyces* were overviewed through comparative genome analysis using pan-genome model. Total 17 *Streptomyces* genomes sequenced completely so far were used for this analysis, cataloging ortholog groups of genes based on the homology. From this, the core genome that conserved in all of analyzed strains was identified. This analysis provides up-to-date information of genomic diversity and core components of *Streptomyces* genome that gives an insight into comprehensive understanding of this genus.

In chapter 4, systematic analysis of transcriptional regulation in *Streptomyces coelicolor* was performed using ChIP-seq technique. First, tandem epitope tagging system was developed for chromatin immunoprecipitation, and was applied to elucidation of the regulatory network of NdgR, a common regulatory protein in *Streptomyces*. The physiological roles of NdgR were further suggested using network motif theory.

In chapter 5, the functional organization and dynamic expression of the genome element in *Streptomyces coelicolor* was dissected using integration of multiple genome-wide data. To obtain the data, TSS-seq and RNA-seq were employed under various growth phases and conditions. Particularly, genomic architecture and expression profiles of secondary metabolite genes were investigated precisely

through integration of those data with unpublished ribosome profiling data that represents translation level using abundance of ribosome protected fragment of mRNA.

The methodology employed in this study would be applied to various systematic studies of *Streptomyces*. Furthermore, the information in this thesis, expanding our knowledge about *Streptomyces* genome, will provide essential database to engineering genetic circuits for the antibiotics synthesis in *Streptomyces*.

Chapter 2. Materials and methods

2.1 Bacterial strains and culture conditions

All strains used are *Streptomyces coelicolor* A3 (2) M145 and *Escherichia coli* K-12 MG1655 and its derivatives. For the ChIP experiments, 6×myc-tagged NdgR was grown on solid minimal media composed of 0.05% (w/w) K₂HPO₄, 0.02% MgSO₄•7H₂O, 0.001% FeSO₄•7H₂O, 2.2% agar, 0.05% L-asparagine, and 1% N-acetylglucosamine. For sensitivity testing and gene expression analysis, spores of *S. coelicolor* A3(2) M145 and an *ndgR* knockout strain, previously constructed using PCR targeting method (Yang et al. 2009b; Kim et al. 2012b), were grown in R5-medium composed of 10.3% sucrose, 0.025% K₂SO₄, 1.01% MgCl₂•6H₂O, 1% glucose, 0.01% Difco casamino acids, 0.2% trace element solution composed of 0.004% ZnCl₂, 0.02% FeCl₃•6H₂O, 0.001% CuCl₂•2H₂O, 0.001% MnCl₂•4H₂O, 0.001% mg Na₂B₄O₇•10H₂O, 0.001% mg (NH₄)₆Mo₇O₂₄•4H₂O in 1 l of deionized water, 0.5% yeast extract, 0.57% TES buffer, and 0.7% (v/v) 1 N NaOH in 1 l of distilled water.

For RNA-seq analysis, Glycerol stock of *S. coelicolor* A3(2) M145 spore was inoculated into R5- complex medium and cultured till OD_{450nm} ~ 0.6. The cultured cell was diluted to 1:100 with fresh R5- medium and grown at 30°C. *S. coelicolor* was harvested in early exponential phase (12 h), mid-exponential phase (16 h), late exponential phase (20 h), and stationary phase (36 h). To prepare Ribo-seq samples, thiostrepton, a translation elongation inhibitor, was added to cell culture to a final concentration of 20 µM and incubated for 5 min at 30°C before harvest. For TSS-seq, cells with 44 conditions were prepared. Cells were harvested at 4 time points in liquid R5- cultures and 3 time points in solid R5- cultures. Cells grown with 32 different nutrient combination (C-sources: glucose, N-acetylglucosamine, glycerol

and maltose; N-sources: ammonium, asparagine, glutamine, serine, leucine, histidine, phenylalanine and casamino acids) or exposed to 5 different stress conditions (0.5 M sodium chloride, 1 % ethanol, 42°C heat shock, 12°C cold shock and 0.01 % SDS) for 1 hour in liquid minimal media were harvested.

2.2 DNA manipulations

2.2.1 Construction of template plasmids for tandem myc tagging

Oligonucleotide encoding tandem myc sequences were chemically synthesized. In accordance with codon usage, the myc sequence (Glu-Gln-Lys-Leu-Ile-Ser-Glu-Glu-Asp-Leu) was optimized for the gene expression in *S. coelicolor* (GAG CAG AAG CTG ATC AGC GAG GAG GAC CTG). PCR was performed using primers carrying EcoRI and BamHI restriction sites in a final volume of 100 µL containing 2.5 U LA Taq DNA polymerase (TAKARA), 50 µl GC buffer I, 35 µl DDW, 8 µl (50 µM) dNTP, 5 µl DMSO and 1 µl (50 ng) plasmid containing the tandem myc tag sequence as a template. PCR amplification conditions were 30 cycles with 30 s denaturation at 94°C, 30 s annealing at 60°C, and 30 s extension at 72°C. The PCR product was then cloned into pUC18 and confirmed by DNA sequencing. A gene cassette, which contained the flanking FRT sites and apramycin resistance gene, was amplified from pIJ773 (Gust et al. 2004) using primers carrying oligonucleotide extensions with BamHI and HindIII restriction sites. The gene cassette was then ligated into the pUC18:(n)-myc to obtain the pJN1 plasmid series.

2.2.2 Tandem epitope tagging to *Streptomyces coelicolor* transcription factors

Linear DNA fragments were amplified using pairs of primers which were 59-bp in

length with 39-bp homology extensions overlapping upstream and downstream from stop codon of target genes and 20-bp priming sequences from pJN1 template plasmid series. Each PCR product was purified, digested with DpnI, repurified, and then electroporated into *E. coli* strain harboring pIJ790 (which expresses the λ Red recombination system under the control of an inducible promoter) and *S. coelicolor* cosmid (which contains a genomic region of interest). The cells were then incubated at 37°C for 1 h in 1mL of LB and spread onto LB-agar medium supplemented with apramycin. The myc-inserted cosmid was transported into the methylation-deficient *E. coli* strain ET12567 containing the RP4 derivative pUZ8002, and then transferred to *S. coelicolor* M145 by conjugation (Flett et al. 1997). Single-crossover exconjugants were selected on MS containing kanamycin and nalidixic acid, to obtain transconjugants. The genomic DNA was then isolated and plasmid integration was confirmed by PCR with the primers of 300bp upstream and downstream from stop codon.

2.3 Chromatin immunoprecipitation

Fifty-milliliter cultures of cells harboring 6×myc-fused transcription factors were grown in R5- complex media at 37°C. The cells were cross-linked by 1% formaldehyde at room temperature for 30 min. Following the quenching of unused formaldehyde with 125 mM glycine at room temperature for 5 min, the cross-linked cells were harvested by centrifugation and washed three times with 50 mL ice-cold Tris-buffered saline (Sigma, St. Louis, MO, USA). The washed cells were resuspended in 1.5 mL lysis buffer composed of 10 mM Tris-HCl (pH 7.5), 100 mM NaCl, 1 mM EDTA, protease inhibitor cocktail (Sigma), and 1 kU lysozyme

(EPICENTRE, Madison, WI, USA). The cells were incubated at room temperature for 30 min and then treated with 2 mL 2× IP buffer (100 mM Tris-HCl, pH 7.5, 200 mM NaCl, 1 mM EDTA, 2% Triton® X-100). The lysate was then sonicated eight times for 20 s each in an ice bath to fragment the chromatin complexes. The range of the DNA size resulting from the sonication procedure was 300–1000 bp, and the average DNA size was 500 bp. Cell debris was removed by centrifugation at 37,000×g for 10 min at 4°C, and the resulting supernatant was used as cell extract for the immunoprecipitation. To immunoprecipitate the protein–DNA complex, 3 µg of anti-c-myc antibody (9E10, Santa Cruz Biotech) were then added into the cell extract, respectively. For the nonspecific control (mock-IP), 2 µg of normal mouse IgG (Upstate) was added into the cell extract. They were then incubated overnight at 4°C, and 50 µL of the Dynabeads Pan Mouse IgG beads (Invitrogen) was added into the mixture. After 5 h of incubation at 4°C, the beads were washed twice with the IP buffer (50 mM Tris-HCl at pH 7.5, 140 mM NaCl, 1 mM EDTA, and 1% (v/v) Triton X-100), once with the wash buffer I (50 mM Tris-HCl at pH 7.5, 500 mM NaCl, 1% (v/v) Triton X-100, and 1 mM EDTA), once with wash buffer II (10 mM Tris-HCl buffer at pH 8.0, 250 mM LiCl, 1% (v/v) Triton X-100, and 1 mM EDTA), and once with TE buffer (10 mM Tris-HCl at pH 8.0, 1 mM EDTA) in order. After removing the TE buffer, the beads were resuspended in 200 µL of elution buffer (50 mM Tris-HCl at pH 8.0, 10 mM EDTA, and 1% SDS) and incubated overnight at 65°C for reverse cross-linking. After reversal of the cross-links, RNAs were removed by incubation with 200 µL of TE buffer with 1 µL of RNaseA (QIAGEN) for 2 h at 37°C. Proteins in the DNA sample were then removed by incubation with 4 µL of proteinase K solution (Invitrogen) for 2 h at 55°C. The sample was then purified with

a PCR purification kit (MACHEREY-NAGEL). To measure the enrichment of the protein-binding targets in the DNA samples, 1 μ L of IP or mock-IP DNA was used to carry out gene specific real-time qPCR with the specific primers to the promoter regions. The *hrdB*, housekeeping sigma factor, promoter sequence was used as a negative control in all cases. All real-time qPCR reactions were done in triplicate. The samples were cycled at 94°C for 15 s, 60°C for 30 s and 72°C for 30 s (total 40 cycles) on a iQ5 (Bio-Rad). The threshold cycle values were calculated automatically by the iCycler iQ optical system software (Bio-Rad Laboratories).

2.4 RNA extraction

The cells were resuspended in 500 μ l lysis buffer (20 mM Tris-Cl pH 7.4, 140 mM NaCl, 5 mM MgCl₂, 1% Triton X-100). Before the resuspension, thiostrepton-treated cells of Ribo-seq samples were rinsed with 5 ml polysome buffer (20 mM Tris pH 7.4, 140 mM KCl, 5 mM MgCl₂, 20 μ M thiostrepton) then the cells were resuspended in lysis buffer as described above. Resuspended cells were dripped into liquid Nitrogen then grinded with mortar. The powdered cells were thawed and centrifuged at 4°C, 3000 g for 5 min to remove cell debris. The supernatant was recovered and centrifuged at 16,000 g for 10 min. The clarified supernatant was recovered. For RNA-seq and TSS-seq samples, total RNA was isolated using miRNeasy Mini kit (Qiagen) according to manufacturer's instruction.

2.5 Differential RNA sequencing (TSS-seq)

Total RNA samples of different culture conditions were isolated as described above. RNAs from liquid R5- complex medium, solid R5- complex plate, liquid minimal

medium, and liquid minimal medium with various stressed conditions described above were collected 2.5 µg each to make 10 µg of total RNA. Genomic DNA was removed by using DNA-free™ Kit (Ambion) following the manufacturer's instructions. To enrich mRNA from the isolated total RNA samples, ribosomal RNA was removed by using Ribo-Zero™ rRNA Removal Kit for Meta-bacteria (Epicentre) according to the manufacturer's instructions. rRNA removed RNAs were verified with the Experion™ system (Bio-Rad). Total mRNA was split into two samples for two different libraries: the library of primary transcriptome and the library of whole transcriptome. 1 U of Terminator™ 5'-Phosphate-Dependent Exonuclease (TEX, Epicentre) was treated to one of the samples to enrich primary transcripts which have triphosphate at their 5' ends. 2 µl of 10x Terminator™ Reaction Buffer A (Epicentre) and 0.5 µl of RNaseOUT™ (40 U/µl, Invitrogen) were added to them and incubated at 30°C for 1 hour. The reaction was terminated by adding 1 µl of 100 mM EDTA pH 8.0. Enriched mRNA was purified by phenol-chloroform extraction and ethanol precipitation. Hereafter, the TEX-treated sample (TEX+) and none treated sample (TEX-) were proceeded together. To ligate 5' RNA adaptor, the triphosphate of 5' end of mRNA was converted to monophosphate by treating 20 U of RNA 5' Polyphosphatase (Epicentre). 2 µl of 10x RNA 5' Polyphosphatase Reaction buffer (Epicentre) and 0.5 µl of RNaseOUT™ (40 U/µl, Invitrogen) were treated together and incubated at 37°C for 1 hour. Then the mRNA was purified by phenol-chloroform extraction and ethanol precipitation. 5 µM of 5' RNA adaptor (GUUCAGAGUUCUACAGUCCGACGAUC) was added to the purified mRNA with 4 µl of T4 RNA Ligase (5 U/µl, Epicentre), 2 µl of 10x T4 RNA Ligase buffer (Epicentre), 2 µl of 10 mM ATP, and 0.5 µl of RNaseOUT™ (40 U/µl, Invitrogen).

The ligation reaction was proceeded by incubation at 37°C for 3 hours. Then cDNA was synthesized from adaptor ligated RNA using random 3' overhanging primer (N9) (GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNNN). The primer-RNA mixture was incubated at 70°C for 10 min then at 25°C for 10 min. And the following was added to the reaction: 6 µl of 10x RT buffer (Invitrogen), 6 µl of 100 mM DTT, 3 µl of 10 mM dNTP mix, 1 µl of actinomycin D (1 mg/ml), 0.75 µl of RNaseOUT™ (40 U/µl, Invitrogen), and 3 µl of SuperScript III Reverse Transcriptase (200 U/µl, Invitrogen). The mixture was incubated 10 min at 25°C, 1 hour at 37°C, 1 hour at 42°C, and 15 min at 70°C, sequentially. The reaction was then chilled to 4°C. To remove residual RNAs, reverse transcribed product was incubated at 65°C for 30 min with 20 µl of 1 N NaOH. Then 20 µl of 1 N HCl was added to the reaction for neutralization. Synthesized cDNA was purified using QIAquick PCR Purification Kit (Qiagen) according to the manufacturer's instructions. The cDNA was purified again by ethanol precipitation. Purified cDNA was selected at a size range between 100 bp and 350 bp on a 2% agarose gel by Pippin Prep (Sage Science). Size selected DNA was purified by ethanol precipitation. Then the purified sequencing library was amplified by PCR with indexed primers for Illumina sequencing platform. The amplification was monitored on a CFX96™ Real-Time PCR Detection System (Bio-Rad) and stopped at the beginning of the saturation point. After enriched library was purified by ethanol precipitation, the library of size between 150 bp and 400 bp was extracted on a 2% agarose gel by Pippin Prep (Sage Science). Size selected library was purified by ethanol precipitation. A second PCR amplification was carried out with a few PCR cycles to produce enough amount of library for Illumina sequencing. The final amplified

library was purified by ethanol precipitation and those of size 150 bp to 400 bp were extracted from 2% agarose gel after electrophoresis. The final library was then purified using MinElute Gel Extraction Kit (Qiagen) and quantified with Qubit 2.0 fluorometer (Invitrogen).

2.6 Strand-specific RNA sequencing (RNA-seq)

Total RNA samples were isolated as described above. To remove genomic DNA, the isolated RNA were incubated at 37°C for 1 hour with 4 U of rDNase I (Ambion) and 5 µl of 10x DNase I buffer (Ambion). The DNA-free RNA was purified by phenol-chloroform extraction and ethanol precipitation. Ribosomal RNA was removed by using Ribo-ZeroTM rRNA Removal Kit for Meta-bacteria (Epicentre) according to the manufacturer's instructions. rRNA removed RNAs were verified with Agilent 2200 TapeStation system (Agilent Technologies). The 200 ng of mRNA was then fragmented by incubation at 70°C for 5 min with 10x Fragmentation buffer (Ambion). The reaction was terminated by adding 1 µl of Stop solution (Ambion) and the fragmented mRNA was purified by ethanol precipitation. For first strand cDNA synthesis, 3 µg of Random primers (Invitrogen) were added to the fragmented mRNA, and they were denatured by incubation at 65°C for 5 min. Then the following was added to the reaction: 2 µl of 10x RT buffer (Invitrogen), 1 µl of 10 mM dNTP mix, 4 µl of 25 mM MgCl₂, 2 µl of 100 mM DTT, 1 µl of SuperScript III Reverse Transcriptase (200 U/µl, Invitrogen), and 1 µl of RNaseOUTTM (40 U/µl, Invitrogen). The mixture was incubated 10 min at 25°C for annealing then 50min at 50°C for reverse transcription. The reaction was terminated by incubation at 85°C for 5 min. Synthesized first strand cDNA was purified by using Agencourt AMPure XP beads

(Beckman Coulter). Following mixture was added to the purified cDNA for second strand synthesis: 1 µl of 10x RT buffer (Invitrogen), 0.5 µl of 25 mM MgCl₂, 1 µl of 100 mM DTT, 2 µl of 10 mM mixture of each dATP, dGTP, dCTP, dUTP, 15 µl of 5x second-strand buffer (Invitrogen), 5 µl of *Escherichia coli* DNA polymerase (10 U/µl, Invitrogen), 1 µl of *E.coli* DNA ligase (10 U/µl, Invitrogen), and 1 µl of *E.coli* RNase H (2 U/µl, Invitrogen). The mixture was incubated at 16°C for 2 hours, and synthesized cDNA was purified by using Agencourt AMPure XP beads (Beckman Coulter). The library for Illumina sequencing was constructed using TruSeq™ DNA Sample Prep Kit (Illumina) according to the manufacturer's instructions. Briefly, the synthesized cDNA was end-repaired and 3' ends of the blunt fragments were adenylated for the adapter ligation. The adenylated DNA fragments were ligated with Illumina adapters. A fraction of the adapter-ligated DNA between 180 bp and 380 bp was size-selected from 2% Agarose gel after electrophoresis. Size-selected DNA was purified by using MinElute Gel Extraction Kit (Qiagen) according to manufacturer's instructions and eluted in 1x TE buffer with low EDTA (10 mM Tris-HCl pH 8.0, 0.1 mM EDTA) for the following enzyme reaction. For degradation of the second strand which contains dUTP instead of dTTP, 1 U of USER enzyme (NEB) was treated to the purified DNA and incubated at 37°C for 15 min. After the 5 min incubation at 95°C for enzyme inactivation, the library was enriched by PCR. The amplification was monitored on a CFX96™ Real-Time PCR Detection System (Bio-Rad) and stopped at the beginning of the saturation point. The final amplified library was purified by using Agencourt AMPure XP beads (Beckman Coulter) and quantified with Qubit 2.0 fluorometer (Invitrogen).

2.7 NGS sequencing

The resulting library was loaded onto a flow-cell and sequenced using MiseqTM v.2 instrument. The 50 bp read recipe was used for RNA-seq and Ribo-seq libraries, and the 150 bp read recipe was used for TSS-seq libraries according to the manufacturer's instructions.

2.8 Western blot analysis

Each sample was subjected to electrophoresis in a SDS-10% polyacrylamide gel, and the resolved proteins were electrotransferred to a HybondTM-ECL membrane (Amersham Biosciences, Piscataway, NJ, USA). The ECLTM Western detection kit (Amersham Biosciences), mouse monoclonal 9E10 antibody, and horseradish peroxidase (HRP)-conjugated sheep anti-mouse IgG (Amersham Biosciences) were used to detect the tandem-myc tagged proteins. Fifty milliliters BCA protein assay kit (Pierce, Rockford, IL, USA) were used to quantify the amount of proteins.

2.9 Bioinformatic analysis

2.9.1 Pan-genome analysis

For the pan-genome computation of *Streptomyces* species, PGAP v1.12 was used (Zhao et al. 2012). Ortholog clusters were organized using open reading frame (ORF) contents of each genome with GF (Gene Family) method. Then, pan-genome and core genome profile was built. Functional enrichment of ortholog clusters performed by PGAP program used cluster of orthologous groups (COG) classification (Tatusov et al. 2001). And the following classification work was performed using in-house script.

2.9.2 ChIP-seq data analysis

Illumina reads were aligned to the *S. coelicolor* reference genome (GenBank: NC_003888) with CLC Genomics Workbench 6.5. The alignment BAM file was analyzed using MACS v1.4 to detect read-enriched regions in the genome (Feng et al. 2012). Enriched regions were selected for further study based on scores greater than 100 using $-10\log_{10}$ (p-value), and fold-enrichment greater than 3. For conserved motif searches, nucleotide sequences (400 bp) centered on each peak were extracted and submitted to the MEME software suite (Bailey et al. 2009). The parameters used were maximum length of 20 nucleotides and zero or one per sequence of each submitted sequence. The FIMO program in the MEME suite was used for searching genome-wide occurrences of the putative motifs derived from MEME.

2.9.3 TSS identification and data analysis

Genomic positions of 5'-end of uniquely aligned TSS-seq reads were considered to potential TSSs. Only TSSs present in both TEX+ and TEX- libraries were retained. TSSs were then determined as described previously (Rach et al. 2009) followed by manual curation. Briefly, potential TSSs within 100 bp were clustered together. And they were sub-clustered with standard deviation <10 . Clusters and sub-clusters with less than three reads were removed. Potential TSS with maximum reads in a sub-cluster was selected as a TSS. If a cluster had several TSSs, standard deviation of two adjacent TSSs was calculated. If calculated standard deviation was less than 10, the one with lower reads was removed. Multiple TSSs or internal TSSs were

considered if they had over 50% of read counts of the maximum TSS. Identified TSSs were compared with previously known TSSs. For searching promoter motifs, the sequences within 10 bp downstream and 50 bp upstream of identified TSSs were extracted from *S. coelicolor* genome. Conserved motif sequences were drawn using Weblogo (Crooks et al. 2004). To identify antisense novel transcripts, genomic positions were divided into 50 bp, and the sum of the normalized read count within each 50-bp window was calculated separately for sense and antisense strands. To remove the artifacts, only the windows which have over 100 of normalized read count were considered, and the windows on antisense strand which have over 40 % of normalized read count of sense strand were retained. Finally, the retained windows were manually curated. Intergenic novel transcripts and novel ORFs were manually identified. All data were visualized using SignalMap™ (Roche NimbleGen, Inc., Madison, WI, USA).

2.9.4 RNA sequencing and ribosome profiling data processing

The linker sequence was trimmed from reads of Ribo-seq library before aligned to the genome. Reads which were shorter than 25 bp after trimming or not contain the linker sequence were discarded. Also, the random 3' overhanging (N9) sequences in reads of TSS-seq library were trimmed. Reads shorter than 25 bp after trimming were discarded. The reads were then aligned to the *S. coelicolor* genome (NC_003888) using CLC Genomics Workbench (CLC bio) with the following parameters: mismatch cost 2, deletion cost 3, insertion cost 3, length fraction 0.9, and similarity fraction 0.9. Only uniquely mapped reads were retained. To validate the coincidence between duplicate data, the expressions of the genes were normalized by RPKM.

Because they showed high correlation in scatter plot (**Figure 2.1**), the expressions of the genes were normalized using DESeq package in R (Anders and Huber 2010). For calculation of fold changes of genes, value of 1 was added to all data to avoid the denominator went to 0. Because the sequences of the first 20 genes and the last 20 genes of *S. coelicolor* are repetitive, those 40 genes were excluded for all analysis.

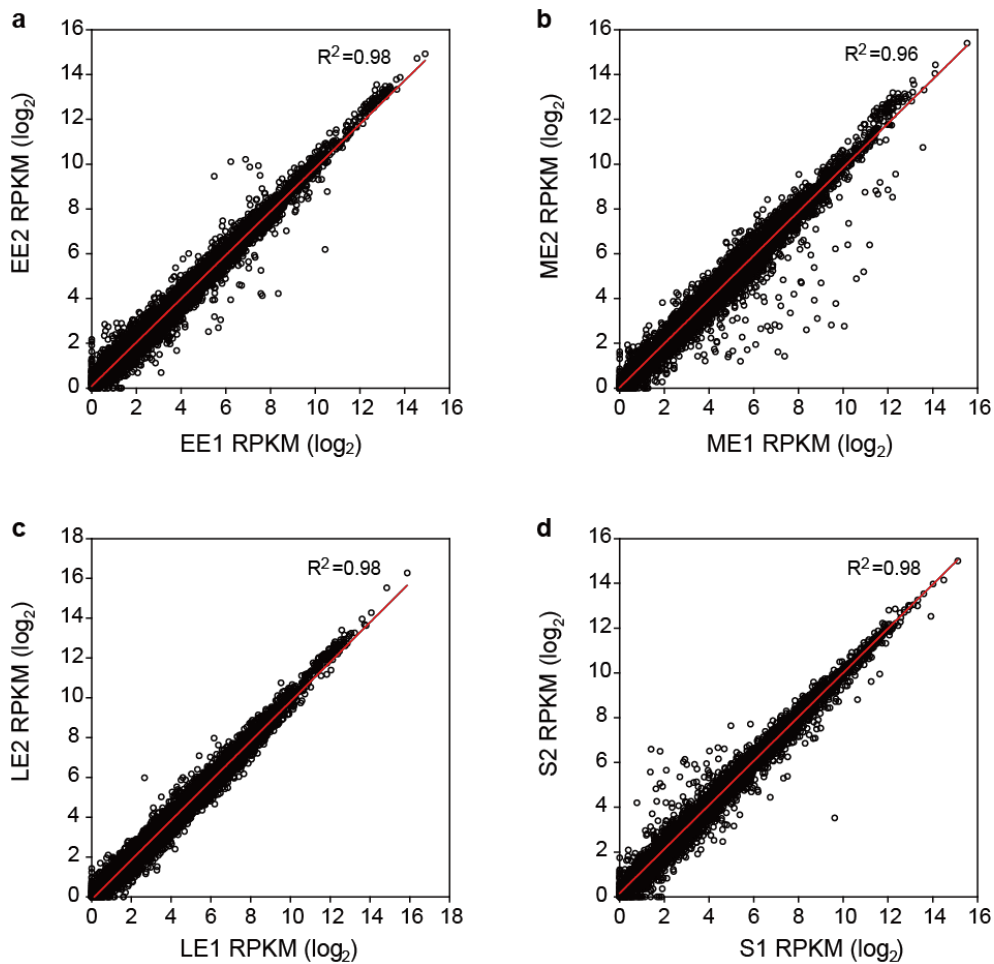


Figure 2.1 Correlation between RNA-seq duplicate data. (a) Early exponential phase (EE) (b) Mid-exponential phase (ME) (c) Late exponential phase (LE) (d) Stationary phase (S).

Chapter 3. Comparative genomics reveals the core and accessory genome of *Streptomyces* species

3.1 The pan-genome of 17 *Streptomyces*

Seventeen completely sequenced *Streptomyces* species genomes available at the NCBI FTP database (<ftp://ftp.ncbi.nih.gov/genomes/Bacteria>) were used in this study. The genomic characteristics of each species are summarized in **Table 3.1**. Thirteen strains contained linear chromosomes, while the other four strains (*Streptomyces* sp. SirexAA E, *S. collinus* TU365, *S. fulvissimus* DSM40593, and *S. vilaceusniger* TU4113) had circular chromosomes. Their genome sizes ranged from 6.3 to 12.7 Mb and had high G+C contents (70.6%–73.3%). The number of predicted coding sequences (CDSs; 5,832–10,022) varied in proportion to their genome sizes. Pan-genome analysis of the 17 *Streptomyces* chromosomes revealed that 34,592 ortholog clusters from 1,129,413 total genes constituted the pan-genome. The size of the *Streptomyces* pan-genome may grow with the number of sequenced strains, and this pan-genome can therefore be considered an open pan-genome (**Figure 3.1**) (Medini et al. 2005). This trend suggests that *Streptomyces* has flexible genome contents, including genes related to their unique secondary metabolism. The core genome consisted of 2,018 ortholog clusters (**Figure 3.2**). This number is smaller than that in a previous report, which described 3,096 gene families derived from five *Streptomyces* strains (Zhou et al. 2012). The ratio of the core genome in each species ranged from 24% to 38% and was negatively correlated with the number of ORFs. Although this number may be decreased when the analyzed genome is added, the number of core genomes would be expected to converge to a constant value according to the slope of exponential decay. The number of dispensable gene families that were conserved in at least two species but not all species was 11,743,

Table 3.1 General genome features of 17 *Streptomyces* species used in this study.

Strain	Length (bp)	G+C (%)	Number		Shape	Accession No.
			of genes	of CDSs		
<i>Streptomyces</i> PAMC26508	7526197	71.1	7054	6969	linear	NC_021055
<i>Streptomyces</i> SirexAA E	7414440	71.7	6648	6362	circular	NC_015953
<i>Streptomyces albus</i> J1074	6841649	73.3	5943	5832	linear	NC_020990
<i>Streptomyces avermitilis</i> MA4680	9025608	70.7	7669	7580	linear	NC_003155
<i>Streptomyces bingchenggensis</i> BCW1	11936683	70.8	10107	10022	linear	NC_016582
<i>Streptomyces cattleya</i> NRRL8057	6283062	72.9	5866	5763	linear	NC_016111
<i>Streptomyces coelicolor</i> A3(2)	8667507	72.1	7910	7769	linear	NC_003888
<i>Streptomyces collinus</i> Tu365	8272925	72.6	7099	7005	circular	NC_021985
<i>Streptomyces davawensis</i> JCM4913	9466619	70.6	8584	8503	linear	NC_020504
<i>Streptomyces flavogriseus</i> ATCC33331	7337497	71.1	6531	6298	linear	NC_016114
<i>Streptomyces fulvissimus</i> DSM40593	7905758	71.5	7027	6925	circular	NC_021177
<i>Streptomyces griseus</i> NBRC13350	8545929	72.2	7224	7136	linear	NC_010572
<i>Streptomyces hygroscopicus jinggangensis</i> 5008	10145833	71.9	8936	8849	linear	NC_017765
<i>Streptomyces rapamycinicus</i> NRRL5491	12700734	71.1	10144	10002	linear	NC_022785
<i>Streptomyces scabiei</i> 87 22	10148695	71.5	8901	8746	linear	NC_013929
<i>Streptomyces venezuelae</i> ATCC10712	8226158	72.5	7536	7448	linear	NC_018750
<i>Streptomyces violaceusniger</i> Tu4113	10657107	71.0	9062	8482	circular	NC_015957

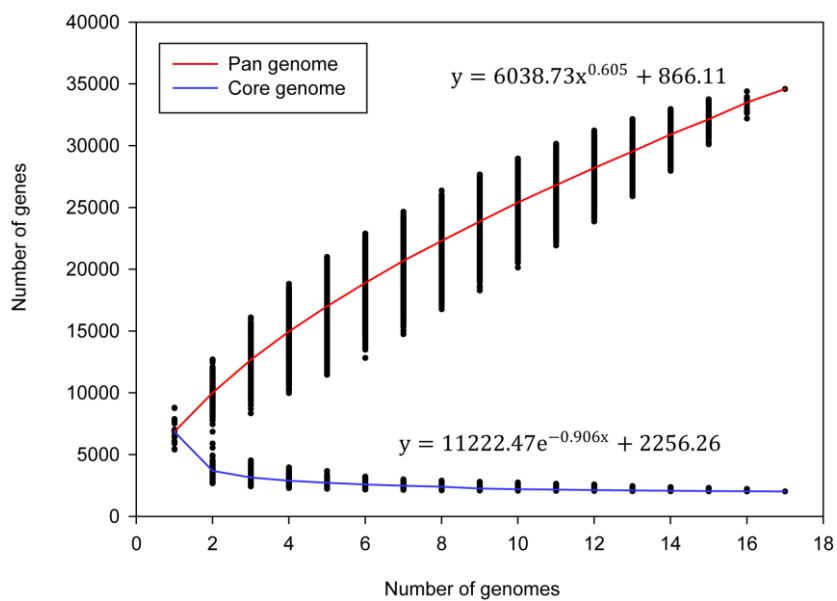


Figure 3.1 Pan-genome and core genome profiles. Accumulated numbers of new genes in the *Streptomyces* pan-genome and core genome are plotted against the number of genomes added. The deduced mathematical function is also reported.

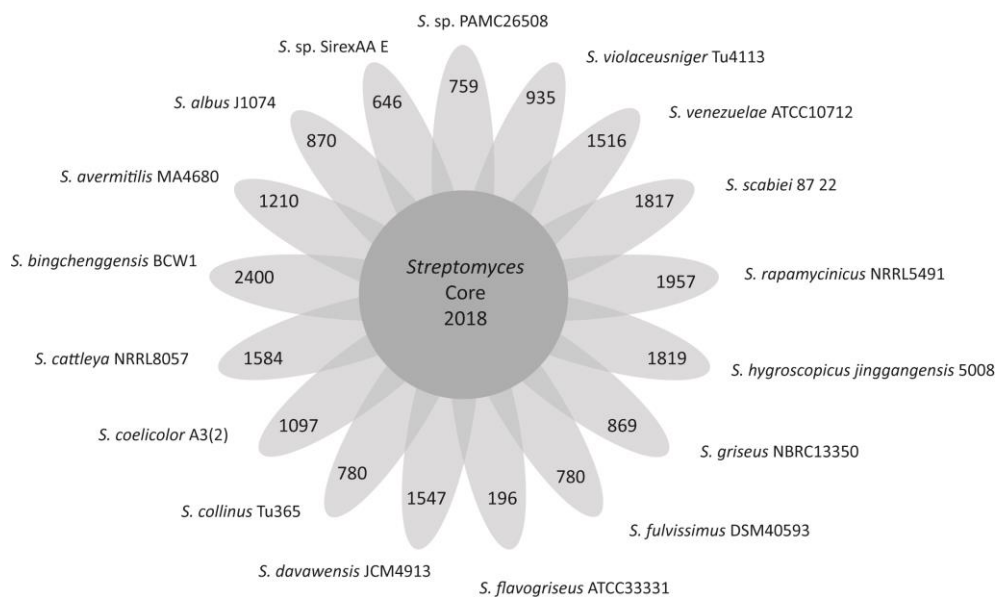


Figure 3.2 Venn diagram of specific genome in each species. The numbers are species-specific gene families in the genome of each species, and the number of core genome is represented in the center.

and the number of ortholog clusters of unique genes that were present in only one strain was 20,831. We called these two groups accessory genomes; these genomes were thought to contribute to the species' diversity and generally provide functions that were not essential to viability. However, these genes may have conferred a selective advantage to *Streptomyces*, such as niche adaptation.

3.2 Functional distribution of ortholog clusters

Next, we examined the functional classifications of ortholog clusters using the COG database (**Table 3.2**). The most abundant category was transcription, which included 1,945 gene families. Various transcriptional regulators and sigma factors were identified in the transcription category. Because *Streptomyces* is well known for its complex transcriptional regulatory networks (Bibb 2005), these genes are expected to be conserved during the response to external stimulus or for adaptation to various environments.

We then investigated the proportion of each conserved group (core, dispensable, and unique genomes) to determine the numbers of genes in each category (**Figure 3.3**). We found that the ratio of core genomes was high in the translation and nucleotide metabolism categories. However, secondary metabolism, RNA processing and modification, and defense mechanisms had low ratios of core genomes. This result suggested that genes responsible for the basic aspects of the biology of a species and its major phenotypic traits were frequently found in core genomes; however, genes responsible for species diversity, supplementary

Table 3.2 COG distribution of ortholog clusters

COG category	whole	core	dispensible	specific
INFORMATION STORAGE AND PROCESSING				
J: Translation, ribosomal structure and biogenesis	353	132	168	53
A: RNA processing and modification	30	2	8	20
K: Transcription	1945	211	1200	534
L: Replication, recombination and repair	599	93	273	233
B: Chromatin structure and dynamics	2	1	1	0
CELLULAR PROCESSES AND SIGNALING				
D: Cell cycle control, cell division, chromosome partitioning	115	22	58	35
Y: Nuclear structure	0	0	0	0
V: Defense mechanisms	296	26	179	91
T: Signal transduction mechanisms	999	114	555	330
M: Cell wall/membrane/envelope biogenesis	630	83	395	152
N: Cell motility	18	4	8	6
Z: Cytoskeleton	6	0	4	2
W: Extracellular structures	0	0	0	0
U: Intracellular trafficking, secretion, and vesicular transport	96	31	45	20
O: Posttranslational modification, protein turnover, chaperones	313	77	173	63
METABOLISM				
C: Energy production and conversion	744	141	424	179
G: Carbohydrate transport and metabolism	1242	130	845	267
E: Amino acid transport and metabolism	1046	192	642	212
F: Nucleotide transport and metabolism	210	78	89	43
H: Coenzyme transport and metabolism	552	107	308	137
I: Lipid transport and metabolism	723	101	411	211
P: Inorganic ion transport and metabolism	531	63	374	94
Q: Secondary metabolites biosynthesis, transport and catabolism	850	42	536	272
POORLY CHARACTERIZED				
R: General function prediction only	2173	243	1357	573
S: Function unknown	1124	130	733	261
Unclassified	22150	234	4268	17648

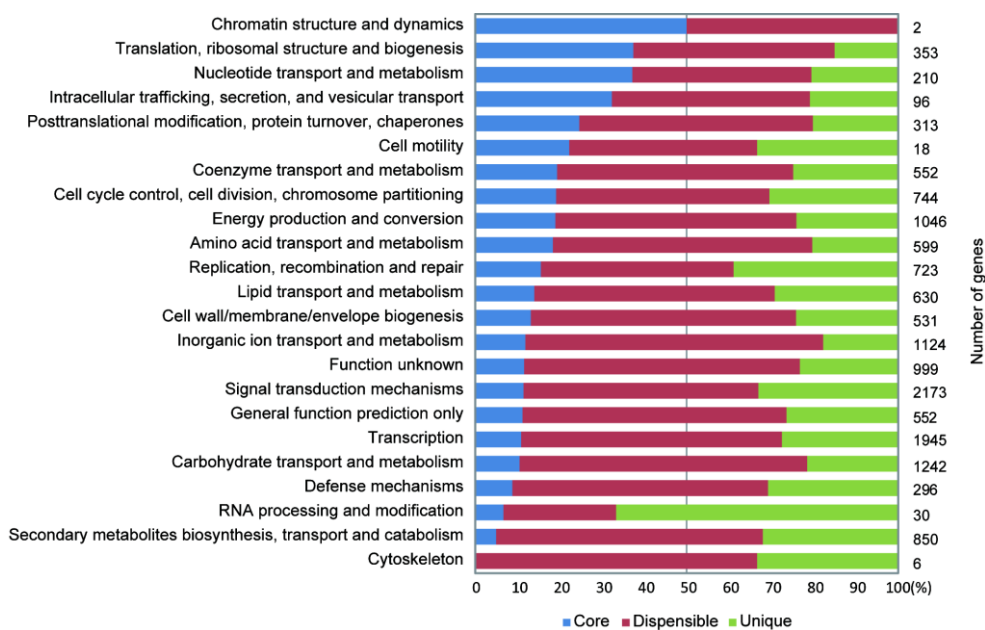


Figure 3.3 Distribution of orthologous genes based on COG category. The bars are sorted by the proportion of core genome in each functional category.

biochemical pathway, and functions that are not essential for bacterial growth were frequently considered accessory genes.

3.3 The core genome in *Streptomyces*

We further investigated the core genome to understand the basic biology of *Streptomyces*. In general, *Streptomyces* species contain a linear chromosome, which has a ‘core region’ that houses the relatively conserved housekeeping genes and two ‘arms’ that contain more divergent and horizontally transferred genes (Dyson 2011). The terminal regions of the chromosomes are highly unstable, and unequal crossing over between the two arms of the chromosome or between one arm of the chromosome and a linear plasmid also occurs frequently, giving rise to gross rearrangements of the chromosome (Dyson 2011). The dynamic nature of the arms is consistent with their high genetic diversity. Therefore, a large part of the terminal region was deleted when the genome-minimized host, which facilitated heterologous expression, was constructed due to the infrequent occurrence of essential genes at the region (Komatsu et al. 2010). We confirmed that there was a high frequency of core genes at the central region of the chromosome in most species, consistent with prior knowledge (**Figure 3.4**).

Next, we examined several important groups of core genes in *S. coelicolor* as a reference strain. First, the complex life cycle of *Streptomyces*, including physiological and morphological differentiation, requires elaborate regulatory systems. This is reflected by the observation that more than 12% of the ORFs in the genome of *S. coelicolor* encode transcriptional regulators. Among these, sigma factors control gene expression at the level of transcription initiation; these factors

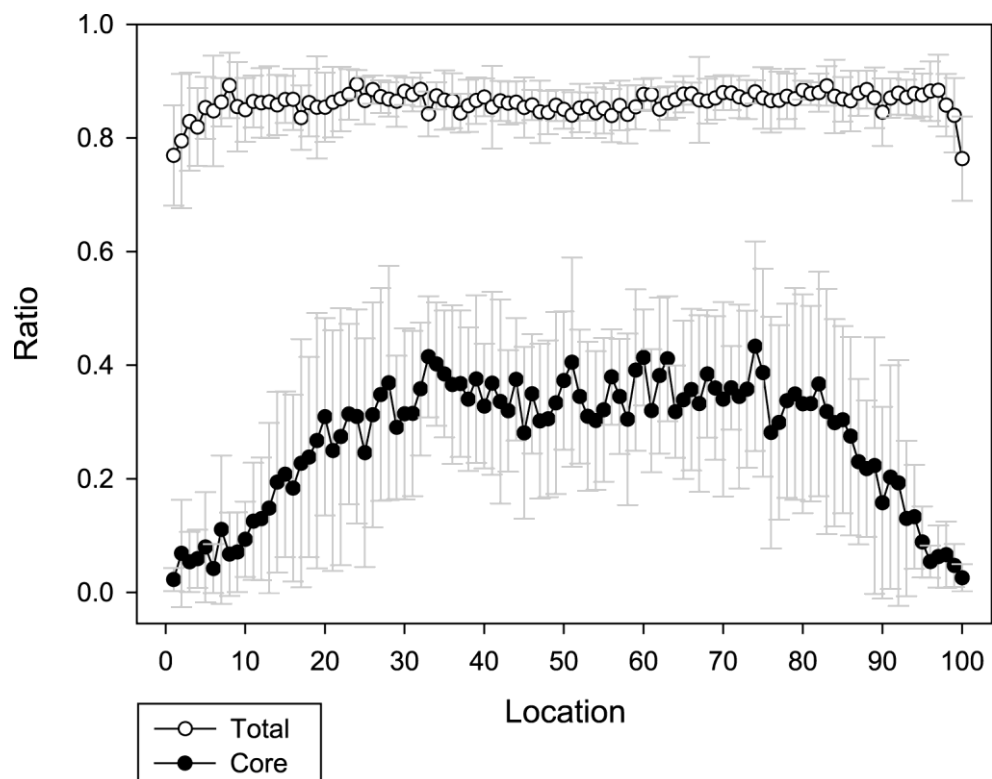


Figure 3.4 Proportion of the core genome according to the location in linear chromosomes. All genomes are normalized to the same size and divided into 100 sections. The plot represents the average ratio of the length of the total and core genes to each section. Error bars indicates standard deviation of the ratio in each section.

have been shown to be important for stress responses and differentiation. Elucidation of core sigma factors in *Streptomyces* therefore facilitates our understanding of the conservation of transcriptional regulatory networks in *Streptomyces* species. We found that 25 out of 65 sigma factors encoded in the *S. coelicolor* genome were included in 15 core ortholog clusters (**Table 3.3**). Major housekeeping sigma factors, such as *hrdA*, *hrdB*, *hrdC*, and *hrdD*, were clustered in an identical gene family, which comprised 3–4 genes in each of the analyzed species, indicating that multiple principle sigma factors were conserved in *Streptomyces*.

In addition, we found that several other previously characterized sigma factors were also conserved. For example, ortholog cluster 4 contained sigma factors such as *sigB*, *sigI*, *sigN*, *sigF*, and *sigH*, which have functions in abiotic stress responses, e.g., responses to osmotic and oxidative stresses (Viollier et al. 2003). Moreover, additional sigma factors related to morphological differentiation, i.e., *bldN*, *sigH*, *sigU*, *sigE*, *sigQ*, and *whiG*, were also included in other ortholog clusters. In particular, *bldN* and *whiG* have been shown to be actively transcribed during aerial mycelium formation and development (Flärdh and Buttner 2009). Some sigma factors with extracytoplasmic functions (ECFs) have also been identified. Among these factors, *sigT* regulates actinorhodin production in response to nitrogen stress (Feng et al. 2011). Taken together, most of these sigma factors regulate multiple genes involved in stress responses, morphological differentiation, and secondary metabolite production. Thus, the conservation of sigma factors suggests the existence of a homologous transcriptional regulatory network in *Streptomyces* species.

Table 3.3 Conserved genes in 17 *Streptomyces* species

Cluster ID	SCO No.	Gene name	Function
Sigma factor			
4	SCO0600	<i>sigB</i>	RNA polymerase sigma factor
4	SCO3068	<i>sigI</i>	RNA polymerase sigma factor
4	SCO4034	<i>sigN</i>	RNA polymerase sigma factor
4	SCO4035	<i>sigF</i>	RNA polymerase sigma factor
4	SCO5243	<i>sigH</i>	RNA polymerase sigma factor
4	SCO7278		RNA polymerase sigma factor
24	SCO0895	<i>hrdC</i>	RNA polymerase sigma factor
24	SCO2465	<i>hrdA</i>	RNA polymerase principal sigma factor
24	SCO3202	<i>hrdD</i>	RNA polymerase principal sigma factor
24	SCO5820	<i>hrdB</i>	RNA polymerase principal sigma factor
107	SCO0942		RNA polymerase sigma factor
107	SCO2954	<i>sigU</i>	RNA polymerase sigma factor
316	SCO4864		ECF sigma factor
316	SCO4866		ECF sigma factor
644	SCO4005		RNA polymerase sigma factor
645	SCO3356	<i>sigE</i>	ECF sigma factor
729	SCO5216	<i>sigR</i>	RNA polymerase sigma factor
882	SCO5147		RNA polymerase sigma factor
956	SCO3613		RNA polymerase sigma factor
1303	SCO4409		RNA polymerase sigma factor
1380	SCO3892	<i>sigT</i>	RNA polymerase sigma factor
1743	SCO3323	<i>bldN</i>	RNA polymerase sigma factor
1910	SCO5621	<i>whiG</i>	RNA polymerase sigma factor
2236	SCO4908	<i>sigQ</i>	RNA polymerase sigma factor
2282	SCO4769		ECF sigma factor
Cell division			
135	SCO3846		FtsW/RodA/SpoVE family cell cycle protein
135	SCO5302		integral membrane cell-cycle protein
350	SCO7046		camphor resistance protein
389	SCO2611	<i>mreB</i>	rod shape-determining protein
768	SCO3560		ATP-binding protein
877	SCO2082	<i>ftsZ</i>	cell division protein
1052	SCO1416	<i>sffA</i>	hypothetical protein
1057	SCO5569		hypothetical protein
1084	SCO1772		partitioning or sporulation protein
1161	SCO2451	<i>mbI</i>	rod shape-determining protein
1194	SCO3406		hypothetical protein
1217	SCO4923		hypothetical protein
1236	SCO2356		hypothetical protein
1325	SCO2969	<i>ftsE</i>	cell division ATP-binding protein
1454	SCO3886	<i>parA</i>	partitioning or sporulation protein
1474	SCO5152		ATP-binding protein
1527	SCO2607	<i>sfr</i>	Sfr protein
1649	SCO5874		hypothetical protein
1707	SCO3557		septum site determining protein
1737	SCO5750	<i>ftsK</i>	ftsK-like protein
1765	SCO2968		cell division protein
1949	SCO3095	<i>divIC</i>	hypothetical protein
2184	SCO5396	<i>filP</i>	cellulose-binding protein

Next, we sought to identify genes related to morphological differentiation, one of the important characteristics of *Streptomyces*, by examination of the cell division category of the COG classification (**Table 3.3**). *Streptomyces* exhibit a complex life cycle that includes spore germination, growth of vegetative mycelium, and formation of aerial hyphae, branched hyphae, and spores (Flärdh and Buttner 2009). Among the core genome, 23 genes in *S. coelicolor* were involved in 22 ortholog clusters, in which many genes that are known to be important for aerial hyphae and spore formation were found. For example, *ftsZ* assembles into a ring structure for the formation of multiple septation (Flärdh and Buttner 2009), and *parA* participates in aerial hyphae and spore formation (Flärdh and Buttner 2009). Several genes involved in chromosome condensation and segregation during cell division were identified in the core genome. *sffA* catalyzes DNA transfer during spore development, and *ftsK* coordinates chromosome segregation during cell division (Flärdh and Buttner 2009). Although *mreB* was conserved in all of the analyzed species, studies of *S. coelicolor* have shown that this gene has little impact on tip extension in vegetative mycelia (Flärdh and Buttner 2009). Additionally, *divIC* plays a role in cell division, but is dispensable for colony formation (Flärdh and Buttner 2009). Some genes identified in this study have not yet been characterized; however, many are expected to have important roles in the differentiation of *Streptomyces*.

Finally, **Table 3.4** shows genes involved in secondary metabolism. Secondary metabolite production is one of the most prominent characteristics of *Streptomyces* species. Biosynthetic genes of secondary metabolites are normally clustered in the

Table 3.4 Conserved genes involved in secondary metabolism in 17 *Streptomyces* species.

Secondary metabolite	SCO No.	Gene name	Function
Coelichelin (1/11)	SCO0489	-	hypothetical protein
5-Hydroxyectoine (4/4)	SCO1864	-	acetyltransferase
	SCO1865	-	diaminobutyrate--2-oxoglutarate aminotransferase
	SCO1866	<i>ectC</i>	L-ectoine synthase
	SCO1867	<i>ectD</i>	hydroxylase
CDA (7/40)	SCO3210	-	2-dehydro-3-deoxyheptonate aldolase
	SCO3211	-	indoleglycerol phosphate synthase
	SCO3212	-	anthranilate phosphoribotransferase
	SCO3213	-	anthranilate synthase component II
	SCO3218	-	hypothetical protein
	SCO3221	-	prephenate dehydrogenase
	SCO3224	-	ABC transporter ATP-binding protein
Siderophore (2/3)	SCO5800	-	hypothetical protein
	SCO5801	-	hypothetical protein
Geosmin (1/1)	SCO6073	-	cyclase
Coelimycin P1 (1/16)	SCO6284	-	decarboxylase
Hopene (11/13)	SCO6759	-	phytoene synthase
	SCO6760	-	phytoene synthase
	SCO6762	-	phytoene dehydrogenase
	SCO6763	-	polyprenyl synthetase
	SCO6764	-	squalene-hopene cyclase
	SCO6765	-	lipoprotein
	SCO6766	-	hypothetical protein
	SCO6767	<i>ispG</i>	4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase
	SCO6768	-	1-deoxy-D-xylulose-5-phosphate synthase
	SCO6769	-	aminotransferase
	SCO6770	-	DNA-binding protein

chromosome. Among the genes related to secondary metabolism, which form 30 clusters (Nett et al. 2009), a total of 27 genes in seven secondary metabolism clusters were identified in the 17 *Streptomyces* species examined in this study. Most of the genes in the 5-hydroxyectoin (4/4), siderophore (2/3), geosmin (1/1), and hopene (11/13) clusters were conserved. 5-Hydroxyectoin is known to have an important role as a compatible solute in response to salt and heat stresses in the *Streptomyces* genus (Bursy et al. 2008). Geosmin, which is responsible for the odor of soil, is also likely to be produced in all of the strains examined in this work (Cane and Watt 2003). Additionally, hopene provides stability at high temperatures and under conditions of extreme acidity due to its rigid ring structure (Poralla et al. 2000). While the secondary metabolites conserved in every species were not specific antibiotics, they acted as protectants in response to stress conditions. Because *Streptomyces* are ubiquitous in various environments that are exposed to stress conditions, such as soil, basic stress response mechanisms are well conserved in the genus.

The amount of bacterial genomic information has been rapidly increased with the development of high-throughput DNA sequencing technologies. In particular, the acquisition and understanding of the genome sequences of *Streptomyces* are important for drug discovery because these organisms are an abundant source of secondary metabolites. In this study, we revealed the conservation of 2,018 and 32,574 genes within core and accessory ortholog clusters, respectively, from 17 completely sequenced *Streptomyces* species using pan-genome analysis. Functional classification of ortholog clusters showed the distribution of ratios of core and accessory genomes. Furthermore, we investigated the functions of the conserved

gene groups, which included sigma factors, cell division pathways, and secondary metabolic pathways. This analysis showed that *Streptomyces* species encoded many common genes involved in the stress response and morphological differentiation. Elucidation of the core genome will provide insights into target selection for genome minimization during the construction of industrial strains or for metabolic engineering. Moreover, this analysis offers a basis for understanding the processes through which information from one strain is transferred to another strain. Integration of genomic information with other -omics studies, such as transcriptomics, proteomics, and metabolomics, will provide an opportunity for understanding more about the functional evolution of *Streptomyces* species.

3.4 Conclusion

The amount of bacterial genomic information has been rapidly increased with the development of high-throughput DNA sequencing technologies. In particular, the acquisition and understanding of the genome sequences of *Streptomyces* are important for drug discovery because these organisms are an abundant source of secondary metabolites. In this study, we revealed the conservation of 2,018 and 32,574 genes within core and accessory ortholog clusters, respectively, from 17 completely sequenced *Streptomyces* species using pan-genome analysis. Functional classification of ortholog clusters showed the distribution of ratios of core and accessory genomes. Furthermore, we investigated the functions of the conserved gene groups, which included sigma factors, cell division pathways, and secondary metabolic pathways. This analysis showed that *Streptomyces* species encoded many common genes involved in the stress response and morphological differentiation. Elucidation of the core genome will provide insights into target selection for genome minimization during the construction of industrial strains or for metabolic engineering. Moreover, this analysis offers a basis for understanding the processes through which information from one strain is transferred to another strain. Integration of genomic information with other -omics studies, such as transcriptomics, proteomics, and metabolomics, will provide an opportunity for understanding more about the functional evolution of *Streptomyces* species.

**Chapter 4. Genome-wide analysis of transcriptional
regulatory network of NdgR in *Streptomyces*
coelicolor using ChIP-seq**

4.1 Construction of PCR-based tandem epitope tagging system for *Streptomyces* genome

The PCR-based tagging strategy applied here starts with amplifying a DNA segment, which begins with the tandem epitope sequence followed by a drug-resistance gene flanked by FRT sites. In addition, the amplifiable segment has homologous sequences to the last portion and to a downstream region of the targeted gene. The precise insertion of the DNA segment into the cosmid containing the target gene is achieved by electroporating the PCR-amplified DNA segment into *E. coli* BW25113/pIJ790 containing the cosmid followed by Red-mediated recombination. The epitope-inserted cosmid is then transported into the methylation-deficient *E. coli* (ET12567/pUZ8002) and transferred to *S. coelicolor* M145 by conjugation (Gust et al. 2003). To provide versatile PCR amplification of the DNA segment, we constructed a series of template plasmids encoding 2×, 4×, and 6×myc epitope. Each template plasmid encodes two priming sites, the repeated epitope sequence with a stop codon, and the apramycin resistance gene flanked by directly repeated FRT sites (**Figure 4.1**).

To demonstrate the functionality of the tandem epitope tagging system, we fused the tandem myc tag to several transcription factors of *S. coelicolor*. One of the targets was a quorum receptor protein, ScbR (SCO6265), whose regulatory function has been previously studied (Chatterjee et al. 2011). Although ScbR has been revealed to have global effects on transcription regulation, little is known

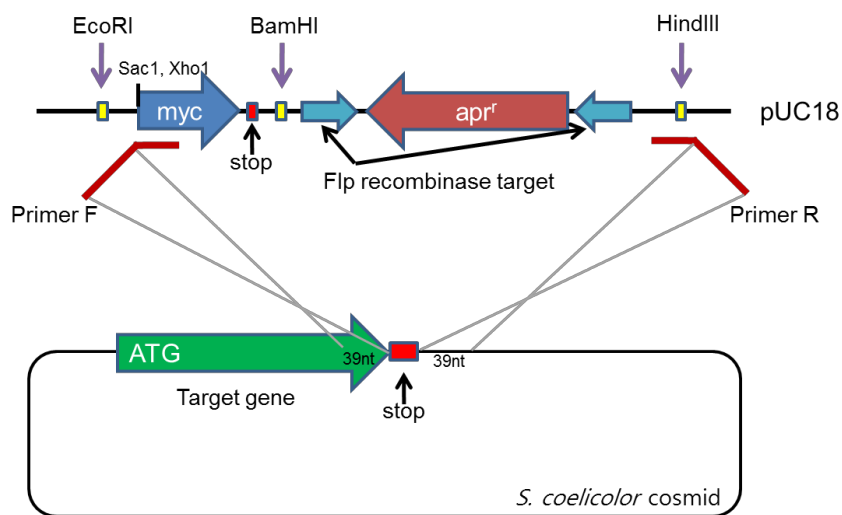


Figure 4.1 PCR-based tandem epitope tagging system for *Streptomyces coelicolor*. A DNA module that involves tandem myc sequence and the antibiotic resistance marker (apr^r) is amplified with primers carrying extensions homologous to the upstream and downstream of the translation stop codon of target gene.

about its direct binding targets except *cpkO* (Takano et al. 2005). Another target was NdgR (SCO5552), an IclR-like regulator which is involved in amino-acid-dependent growth, quorum sensing, and antibiotic production. NdgR showed the binding to intergenic region of *ndgR-leuC* in vitro (Yang et al. 2009b), however none of in vivo measurement has been made to explore regulatory interaction between NdgR and such cis-acting elements. The correct insertion of tandem epitope into the desired genomic loci was first validated by PCR and subsequently by western blot analysis. To this end, genomic DNA was extracted from each tagged strain as a PCR template. PCR was then carried out by using the primer pair of both sides between the start codon and the end of apramycin cassette (**Figure 4.1**). The PCR products which have several restriction sites capable of one-step confirmation were digested to three DNA fragments after XhoI/BamHI treatment (**Figure 4.2A**). While the apramycin resistance gene and open reading frame (ORF) region indicate the same size (the upper and middle bands, respectively), the lowest bands contained DNA sequences encoding the tandem myc epitope and varied in size depending on the number of tag repeats. Western blot analysis revealed that the intensity and the size of tandem epitope-fused transcription factors were proportional to the number of tandem myc copies as expected (**Figure 4.2B**).

4.2 Verification of tagging system using chromatin immunoprecipitation

With confirming the correct integration of the tandem epitope-encoding sequence into the desired locus, the most important aspect of tandem epitope tagging is to maintain the proper function of tandem epitope-fused proteins (Nègre et al. 1991). The major uncertainty in an epitope-tagging strategy is that the fused proteins

potentially lose their in vivo functions. To address the potential drawback, we compared the mRNA expression level of the target genes and the phenotype of tagged strains with those of the wild-type. First, we compared the regulatory function of 6×myc fused ScbR with that of wild-type ScbR. As a quorum receptor protein, ScbR plays an important regulatory role in the onset of antibiotic production in *S. coelicolor* with a signaling molecule, a γ -butyrolactone SCB1 (Takano et al. 2005). In the absence of signaling molecules, ScbR represses transcription of the cryptic type I polyketide synthase gene cluster (*cpk*) by directly binding to promoter region of *cpkO* (SCO6280), the activator of the *cpk* gene cluster (Takano et al. 2005). On the other hand, SCB1 at high concentration binds to ScbR to form a SCB1-ScbR complex, thereby relieving the repression. Thus *cpkO* is constitutively expressed in a *scbR* null mutant irrespective of the presence of SCB1. While the negative regulation of the *cpkO* gene was abolished in the *scbR* null mutant strain, the JN116 strain harboring the 6×myc fused ScbR maintained the negative regulatory effect at the same magnitude as in the wild-type under early exponential growth phase (**Figure 4.3A**). Phenotypically, when liquid cultured cell was transferred to and spread out on a solid R5- medium, the wild-type strain forms spores (**Figure 4.3C**). As in the case with the wild-type strain, the formation of spores was clearly observed from the JN116 strain. However, the deletion of *scbR* gene abolishes the formation of spores. This observation, along with the expression data shown earlier, demonstrates that the ScbR maintained its regulatory functions.

In a previous study, deletion of the *ndgR* gene causes slow cell growth and various phenotypes depending upon which amino acids were used in the minimal media

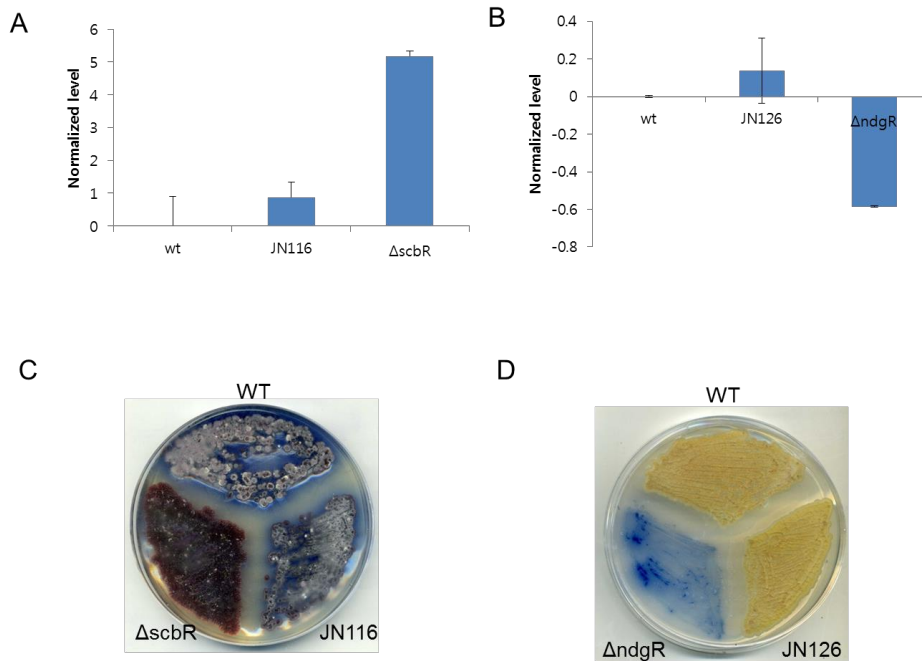


Figure 4.3 Conservation of in vivo function of *scbR*-6X myc and *ndgR*-6X myc tagged strain. (A) CpkO expression in wild-type (M145), *S. coelicolor* JN116, and ScbR null mutant. (B) LeuC expression in wild-type (M145), *S. coelicolor* JN126, and NdgR null mutant. (C) Phenotype on the R5- solid media plate. JN116 produced spores like WT but *scbR* knockout mutant did not have. (D) Phenotype on the GlcNAc/ASN minimal media plate. NdgR knockout mutant showed slow growth rate but produced actinorhodin more than wild-type and JN126. Wild-type and JN126 had similar phenotypes.

(Yang et al. 2009b). In particular, growth of the *ndgR* null mutant in the minimal medium supplemented with N-acetylglucosamine and asparagine increased the production of actinorhodin (ACT) compared to that of the wild-type. However, JN126 strain harboring 6×myc fused NdgR exhibited similar level of ACT production with the wild-type (**Figure 4.3D**). Judging from the ACT production levels, the 6×myc tag does not significantly affect the regulatory role of the NdgR. Based upon the enhancement of ACT production level of the *ndgR* null mutant grown in minimal media containing certain amino acids such as leucine, the regulatory interaction was revealed by electrophoretic mobility shift assay between NdgR and *ndgR-leuC* intergenic region (Yang et al. 2009b). Moreover, expression of the *leuC* gene (SCO5553) in JN126 strain was similar to the wild-type, whereas transcript level of the *leuC* gene was decreased in the *ndgR* null mutant (**Figure 4.3B**). Consequently, it was concluded that both tandem myc-tagged ScbR and NdgR retained their regulatory function of transcription over *cpkO* and *leuC*, respectively.

Tandem epitope fused ScbR and NdgR did not affect in vivo functions of the transcription factors, thus recombinants could be directly used for chromatin immunoprecipitation (ChIP) experiments. This ChIP technique is useful to demonstrate cellular DNA-protein interactions under a diverse set of physiological conditions (Spencer et al. 2003). In ChIP procedure, the cross-linked DNA-protein complex is enriched through immunoprecipitation by the specific antibody against the protein of interest. Also, highly stringent salt and detergent conditions are used to remove nonspecific interactions between cellular proteins and the antibody being used. As a result, affinity and specificity between the protein of interest and the

specific antibody is critical to increase the enrichment yield. As shown in **figure 4.4**, interactions between 6×myc fused protein and the target promoter regions were determined by the quantitative PCR. First, the enrichment level of *cpkO* promoter, known as a target of *scbR*, was 100-fold greater than that of nonspecific ChIP sample (**Figure 4.4A**). Similarly, the normalized ChIP DNA value of *ndgR* promoter region came out to be 8 times greater compared to that of nonspecific sample (**Figure 4.4B**). As a negative control experiment, no differences were observed between specific and nonspecific DNA using a promoter region of a housekeeping sigma factor, *hrdB*. Judging from the enrichment of the known ScbR and NdgR binding sites, it meant that tandem myc tagged proteins maintained their correct interactions with their target regulatory sites.

4.3 Identification of in vivo NdgR binding regions by ChIP-seq

To determine the NdgR-binding regions at the genome scale, we constructed a sequencing library using IP-DNA and performed next-generation sequencing. Sequencing of the library yielded short sequence reads of 36 nucleotides that were uniquely mapped onto the *S. coelicolor* genome (NC_003888). Using the MACS program, 19 NdgR-binding loci were detected with stringent cut-off conditions (p-value <1.0e-10, fold enrichment > 3) (**Table 4.1**). The peaks were distributed across the entire *S. coelicolor* genome (**Figure 4.5A**).

We annotated target genes according to the location of peak summits. If a peak summit was located in ≤ 500 bp upstream or ≤ 100 bp downstream of an

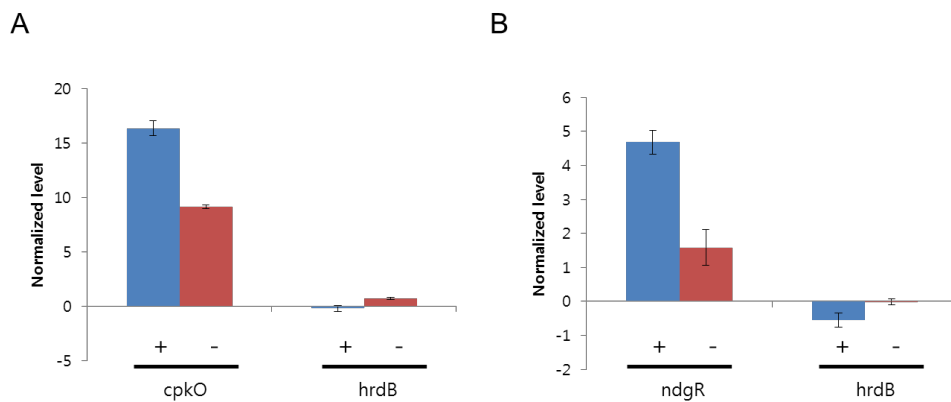


Figure 4.4 Verification of tagging system using chromatin immunoprecipitation (ChIP). Enrichment of promoter regions in specific DNA sample and nonspecific DNA sample was monitored by quantitative real-time PCR (qRT-PCR). (A) Normalized DNA quantity in specific DNA ChIP sample and nonspecific DNA ChIP sample of *scbR*. (B) Normalized DNA quantity in specific DNA ChIP sample and nonspecific DNA ChIP sample of *ndgR*.

Table 4.1 Genome-scale identification of NdgR binding regions.

Start	End	Summit	<i>p</i> -value ^a	Fold enrichment	Motif position	Consensus sequence
1662185	1663375	1662816	197.21	4.41	1662781-1662795	GTCCACCCACCGGAC
1827404	1829415	1828601	768.68	6.04	1828423-1828437	GTCCACCACGCGGAC
1896843	1898738	1897467	276.74	5.04	1897508-1897522	GTCCATCCTGCGGAC
2261363	2262539	2261945	101.39	3.54	2261989-2262003	GTCCACCCTTTGGAC
3161458	3164611	3163260	370.89	4.20	3163207-3163221	GATCACACTCCGGAA
3268432	3272122	3269947	139.50	4.06	3269922-3269936	GTTCACCTCGTGGTC
3702919	3706333	3704715	225.72	4.78	3704846-3704860	TTCCACCTTGATCAC
4578730	4584188	4580680	129.09	5.24	4580694-4580708	TCCCACTCCTTGGAC
4600435	4603056	4601769	162.13	4.23	4601788-4601802	GGTCAGCTCCTGGAC
5625694	5627962	5626280	123.46	3.01	N.A.	N.A.
5741825	5743033	5742489	133.14	3.08	5742420-5742434	GACCACCTCGTGGAC
5862424	5864373	5863498	180.16	4.28	5863483-5863497	GTCCACACCGTGGAC
5880856	5884347	5882679	160.34	3.32	5882573-5882587	GTCCGCCTTGAGGAC
6001446	6003903	6002834	279.72	3.77	6002862-6002876	TCCCACCCATTGAC
6013959	6016845	6015210	329.68	5.54	6015204-6015218	GTCCGCCATGCGGAC
6048531	6053535	6051804	218.57	4.40	6051783-6051797	GTCCAGAACGCCGAC
6059038	6060496	6059807	175.83	4.20	6059774-6059788	GTCCAGCAAGTGGAC
6701783	6704095	6702908	291.94	3.29	6702861-6702875	GTCCACATTTTGAT
7199604	7200985	7200371	425.57	5.00	7200390-7200404	TTTGCCATGTGGAC

a. $-10\log_{10}(p\text{-value})$.

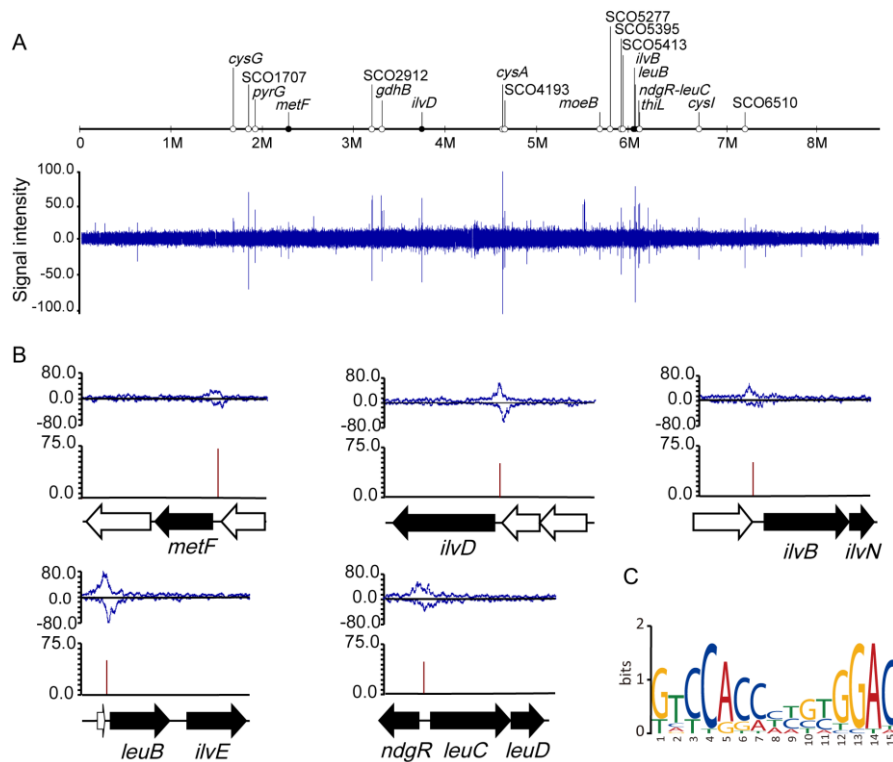


Figure 4.5 Genome-wide distributions of NdgR binding regions. (A) An overview of NdgR binding profiles across the *S. coelicolor* genome when grown on solid minimal media supplemented with N-acetylglucosamine and L-asparagine. Black and white dots indicate previously known and newly found NdgR binding regions, respectively. (B) Examples of binding profiles of previously known targets of NdgR. Red lines indicate the locations of putative binding motifs derived from FIMO and the values are the scores for the match of a position. Black arrows indicate the target genes within the transcription units that are directly regulated by NdgR. (C) MEME logo representation of the NdgR-DNA binding profile. This motif is present in 18 out of 19 enriched regions identified by ChIP-seq.

annotated start codon, NdgR was considered to regulate the corresponding gene. We annotated target genes according to the location of peak summits. If a peak summit was located in ≤ 500 bp upstream or ≤ 100 bp downstream of an annotated start codon, NdgR was considered to regulate the corresponding gene. When the summit was located in an intragenic region between ≥ 100 bp from the start codon of a relevant gene and ≥ 500 bp upstream from the start codon of a downstream gene, the binding peak was annotated as an intragenic binding. Locations of 15 out of 19 peaks were assigned as the intergenic region upstream of transcription units (<http://biocyc.org/SCO/organism-summary?object=SCO>). Meanwhile, the peak summits of 4 genes (SCO1776, SCO2999, SCO4193, and SCO5277) were located into an intragenic region. With these criteria, we identified 34 genes in the NdgR regulon (**Table 4.2**).

Five genes, *metF* (SCO2103), *ilvD* (SCO3345), *ilvB* (SCO5512), *leuB* (SCO5522), and *ndgR-leuC* intergenic region (SCO5552-SCO5553) were previously annotated as direct regulatory targets of NdgR by in vitro experiments such as electrophoretic mobility shift assay and DNA affinity capture assay (Yang et al. 2009b; Kim et al. 2012b). The results of the current study show highly enriched profiles of these genes in the ChIP-seq data (**Figure 4.5B**). However, an exception was found at the promoter region of *scbR* (SCO6264), which is a gene directly regulated by NdgR (Yang et al. 2009b). This discrepancy might be due to the differences between in vivo and in vitro experimental conditions. In addition, we observed low levels of NdgR binding at the promoters of *metH* (SCO1657) and *leuA* (SCO5559) relative

Table 4.2 The NdgR regulon genes.

SCO No.	Name	Function	Category ^a	Note ^b
SCO1552		rRNA methylase	2.2.11 RNA synthesis, modification, DNA transcript'n	
SCO1553	<i>cysG</i>	Putative uroporphyrin-III methyltransferase	3.2.6 Heme, porphyrin	*
SCO1707		Putative ABC sugar transporter, ATP-binding subunit	1.5.0 Transport/binding proteins	*
SCO1776	<i>pyrG</i>	Putative CTP synthetase	3.3.11 Nucleotide interconversions	*, I
SCO2103	<i>metF</i>	5,10-methylenetetrahydrofolate reductase	3.1.14 Methionine	*
SCO2910	<i>cysM</i>	Cysteine synthase	3.1.6 Cysteine	
SCO2911		Hypothetical protein	0.0.2 Conserved in organism other than <i>Escherichia coli</i>	
SCO2912		Hypothetical protein	0.0.0 Unknown function, no known homologs	*
SCO2999	<i>gdhB</i>	Glutamate dehydrogenase	0.0.2 Conserved in organism other than <i>Escherichia coli</i>	*, I
SCO3345	<i>ilvD</i>	Dihydroxy acid dehydratase	3.1.21 Valine	*
SCO4164	<i>cysA</i>	Putative thiosulfate sulfurtransferase	3.3.19 Sulfur metabolism	*
SCO4165		Hypothetical protein	0.0.2 Conserved in organism other than <i>Escherichia coli</i>	
SCO4193		Putative ATP/GTP-binding membrane protein	4.1.6 Gram +ve membrane	*, I
SCO5178	<i>moeB</i>	Putative sulfurylase	3.2.14 Thiamine	*
SCO5277		Magnesium chelatase	7.0.0 Not classified (included putative assignments)	*, I
SCO5395		Putative ABC transporter ATP-binding subunit	1.5.0 Transport/binding proteins	*
SCO5413		Possible MarR-transcriptional regulator	6.3.7 MarR	*
SCO5512	<i>ilvB</i>	Acetolactate synthase	3.4.3 Carbon compounds	*
SCO5513	<i>ilvN</i>	Acetolactate synthase 3 regulatory subunit	3.1.21 Valine	
SCO5514	<i>ilvC</i>	Acetolactate synthase small subunit	3.1.21 Valine	
SCO5522	<i>leuB</i>	3-isopropylmalate dehydrogenase	3.1.12 Leucine	*
SCO5523	<i>ilvE</i>	Branched-chain amino acid aminotransferase	3.1.21 Valine	
SCO5552	<i>ndgR</i>	Putative regulator	6.5.0 Others	*, D
SCO5553	<i>leuC</i>	Isopropylmalate isomerase large subunit	3.1.12 Leucine	*, D
SCO5554	<i>leuD</i>	Isopropylmalate isomerase small subunit	3.1.12 Leucine	
SCO5562	<i>thiL</i>	Thiamin monophosphate kinase	3.2.14 Thiamine	*
SCO5563	<i>thiD</i>	Phosphomethylpyrimidine kinase	3.3.14 Thiamine	
SCO6097	<i>cysN</i>	Sulfate adenylyltransferase subunit 1	3.3.19 Sulfur metabolism	
SCO6098	<i>cysD</i>	Sulfate adenylyltransferase subunit 2	3.3.19 Sulfur metabolism	
SCO6099	<i>cysC</i>	Adenylylsulphate kinase	3.3.19 Sulfur metabolism	
SCO6100	<i>cysH</i>	Phosphoadenosine phosphosulfate reductase	3.3.19 Sulfur metabolism	
SCO6101		Hypothetical protein	0.0.0 Unknown function, no known homologs	
SCO6102	<i>cysI</i>	Putative nitrite/sulfite reductase	3.5.2 Anaerobic respiration	*
SCO6510		Conserved hypothetical protein	0.0.2 Conserved in organism other than <i>Escherichia coli</i>	*

a. Categories are defined by functional classification of *S. coelicolor* genes in The Sanger Institute database (ftp://ftp.sanger.ac.uk/pub/S_coelicolor/classwise.txt).

b. Genes with direct binding by NdgR are marked with asterisks (*). Binding of the intragenic regions is denoted as I. Binding of the upstream region between two divergent genes is denoted as D.

to the other binding peaks. Taken together, these results show that binding locations of NdgR were successfully detected in vivo.

4.4 Sequence analysis of NdgR-binding regions

Although the putative DNA-binding motifs of NdgR and its orthologs were previously predicted using the known binding sequences of other IclR-type regulators in different strains (Yang et al. 2009b; Kim et al. 2012b; Santamarta et al. 2007; Brune et al. 2007), clear consensus sequences remain undefined. To determine the putative NdgR binding motif from our validated ChIP-seq data, 400 nucleotides surrounding the peak summits of 19 binding regions were analyzed by MEME, a bioinformatics tool that identifies overrepresented motifs in multiple unaligned sequences. A 15-bp imperfect palindromic motif ([GT]T[CT]CAC[CA][CTA][TC][GC][TC]GGAC) was detected with an E-value of 1.3e-005 (**Figure 4.5C**). This motif was present in 18 out of 19 binding loci, and 13 of them were located within 50 bp of each peak summit. Although the sequence is dissimilar to any known motifs of IclR-type regulators, palindromic sequences of 15 bp have been identified as binding sequences for other IclR-type regulators, and are consistent with a helix-turn-helix interaction (Lynda et al. 2008; Pan et al. 1996).

We identified the putative binding motif from previously known targets of NdgR as revealed by in vitro experiments such as electrophoretic mobility shift assay and DNA affinity capture assay (Yang et al. 2009b; Kim et al. 2012b). This motif was detected in upstream regions of *metF*, *ilvB*, and *leuB*, and the intergenic region between NdgR and *leuCD*. The putative motif from genome-wide prediction using FIMO was found in the promoters of known targets, *metH* and *leuA*. This result

further supports that their low intensity in ChIP-seq profiles represents the subtle differences between in vivo and in vitro experimental conditions.

4.5 Functional classification of the NdgR regulon

Genes in the NdgR regulon were further classified into functional categories according to gene classifications defined by The Sanger Institute database (ftp://ftp.sanger.ac.uk/pub/S_coelicolor/classwise.txt) (**Figure 4.6**). The metabolism of small molecule category is highly dominant (62%, 21/34 genes). Among the genes in this category, 43% and 33% were assigned to amino acid biosynthesis (9/21 genes) and central intermediary metabolism (7/21 genes), respectively, and 14% (3/21 genes) were included in biosynthesis of cofactors and carriers.

NdgR directly regulates eight genes in the biosynthetic pathways of branched chain amino acids (BCAAs) (**Figure 4.7A**). For instance, the first step in BCAA biosynthesis is catalyzed by acetohydroxy acid synthase/acetolactate synthase encoded by *ilvBN*. This enzyme catalyzes the condensation of two pyruvate molecules to acetolactate and 2-acetohydroxybutyrate from pyruvate and 2-ketobutyrate. The following reaction is catalyzed by ketol-acid reductoisomerase and dihydroxy-acid dehydratase encoded by *ilvC* and *ilvD*, respectively. The final transamination step, as well as the first step in the degradation pathways, is catalyzed by BCAA aminotransferases encoded by *ilvE*. Leucine is synthesized from α -ketoisovalerate, an intermediate in the valine pathway, through three enzymatic steps. The relevant enzymes are α -isopropylmalate synthase (LeuA), β -isopropylmalate dehydratase (LeuC, LeuD), and β -isopropylmalate dehydrogenase

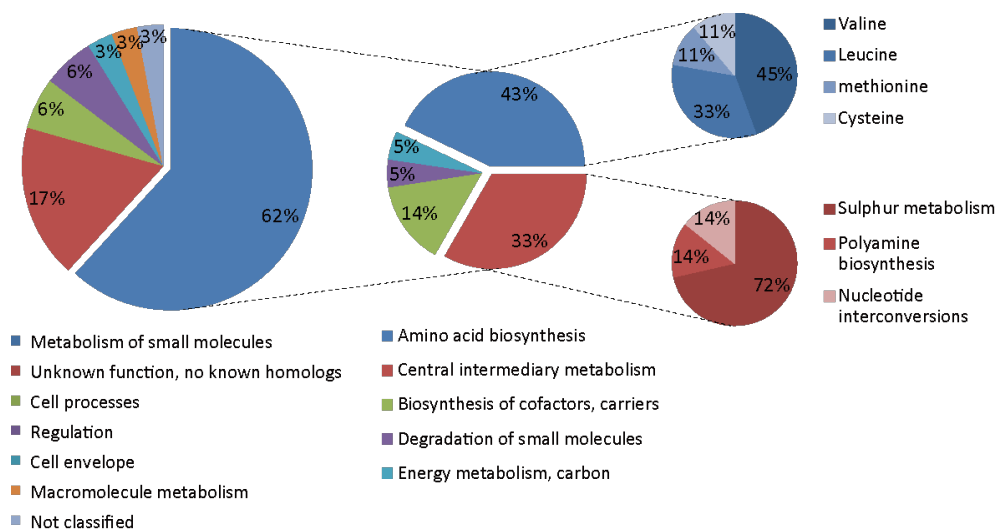


Figure 4.6 Functional classification of genes in the NdgR regulon. Hierarchical functional class is defined by The Sanger Institute database. Genes in this chart are described in Table 2.

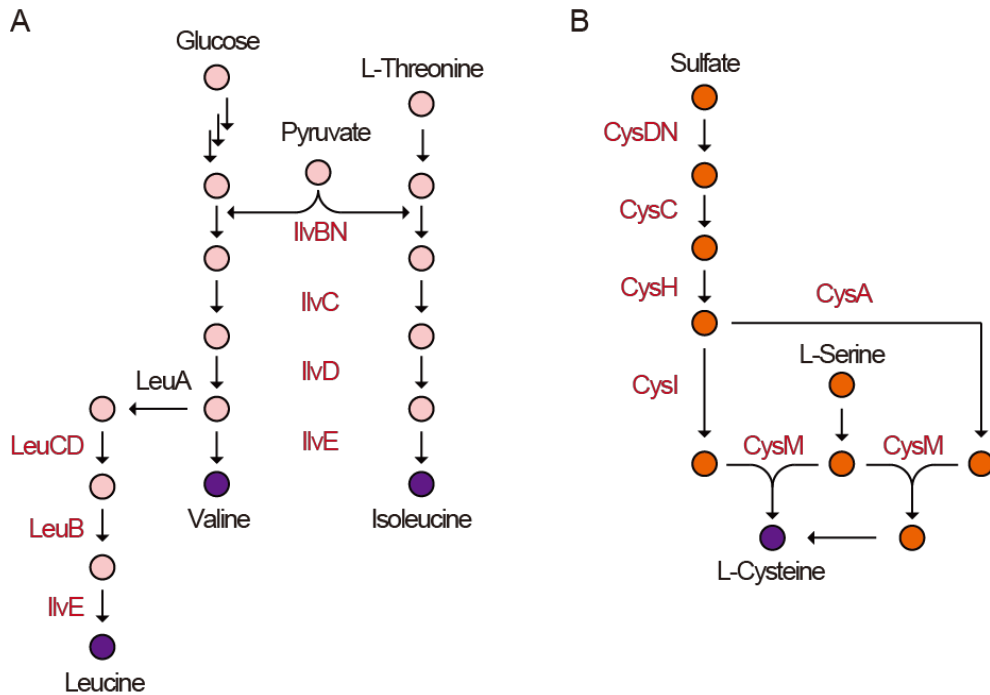


Figure 4.7 Metabolic pathways directly regulated by NdgR. The genes identified by ChIP-seq are depicted by red characters. (A) NdgR directly regulates genes in most steps of the BCAA biosynthesis pathways. Though *leuA* was not annotated as a member of NdgR regulon, the putative motif from genome-wide prediction using FIMO and low binding signal in ChIP-seq data was observed in its upstream region. (B) The sulfur assimilation into the cysteine biosynthesis pathways.

(LeuB). Despite the scattered locations of those genes along the chromosome, NdgR directly bound their upstream regions. Interestingly, the NdgR mutant (BG11) exhibited methionine auxotrophy, but not leucine auxotrophy (Kim et al. 2012b). Notwithstanding its direct binding at the promoters of genes in the BCAA biosynthetic pathway, NdgR was not essential for BCAA biosynthesis under the growth conditions used here. Thus, it is expected that the NdgR plays a role in the fine-tuning of BCAA biosynthesis with the assistance of feedback regulation and translational attenuation common in amino acid biosynthetic pathways in bacteria (Kopecky et al. 1999; Craster et al. 1999; Seliverstov et al. 2005).

Next, we observed that NdgR directly bound the upstream region of three transcription units including seven genes (*cysA*, *cysM*, *cysN*, *cysD*, *cysC*, *cysH* and *cysI*) that are involved in sulfur-assimilation metabolism (**Figure 4.7B**). The pathway of sulfur assimilation into cysteine biosynthesis in *S. coelicolor* has been suggested in previous reports (Lee et al. 2005; Derek et al. 1988). The genes, including *cysN* (SCO6097), *cysD* (SCO6098), and *cysC* (SCO6099), participate in 3'-phosphoadenylyl sulfate (PAPS) formation from sulfate. PAPS reductase encoded by *cysH* (SCO6100) converts PAPS to sulfite. The serial reactions are followed by two pathways that result in thiosulfate production by a thiosulfate sulfurtransferase and sulfide production by a sulfite reductase encoded by *cysA* (SCO4164) and *cysI* (SCO6102), respectively. The two metabolites (thiosulfate and sulfide) are sulfur donors for the sulfur assimilation into O-acetyl-L-serine by cysteine synthase encoded by *cysM* (SCO2910) (Donadio et al. 1990; Fischer et al. 2012). Moreover, NdgR bound to the promoter region of putative siroheme synthase encoded by *cysG* (SCO1553). Siroheme is a prosthetic group that participates in six-

electron reduction reactions catalyzed by both sulfite and nitrite reductases. CysG converts uroporphyrinogen III, which is a precursor of heme and cobalamin (vitamin B12), to siroheme using multifunctional activities such as SAM-dependent methylase, dehydrogenase and ferrochelatase (Stroupe et al. 2003). Biosynthesis of sulfur-containing thiamine was also regulated by NdgR. As the active form of thiamine, thiamine pyrophosphate (TPP) is an essential cofactor for particular metabolic processes such as BCAA biosynthesis. The thiamine category includes thiamine monophosphate kinase and sulfurylase encoded by *thiL* (SCO5562) and *moeB* (SCO5178), respectively. In addition, phosphomethylpyrimidine kinase encoded by *thiD* (SCO5563) in the polyamine biosynthesis category is likely involved in thiamine metabolism.

In *S. clavuligerus*, AreB, an ortholog of NdgR, bound to the upstream region of the pathway-specific regulator of clavulanic acid and cephamycin C, and the *areB* deletion mutant increased their production. Moreover, the *ndgR* mutant showed overproduction of actinorhodin in our experimental conditions. Interestingly, though some effects of secondary metabolites have been observed in the mutants of *ndgR* and its ortholog (Yang et al. 2009b; Santamarta et al. 2007), none of the secondary metabolite genes were detected as NdgR targets. This observation indicates that NdgR indirectly regulates the genes involved in secondary metabolism. Taken together, the data suggest that NdgR mainly regulates primary metabolism of small molecules, especially BCAA and several sulfur-containing molecules.

4.6 Role of NdgR under thiol oxidative stress

To ascertain the physiological role of the IclR family of regulators in the cell,

identification of the interaction between the ligand and the substrate-recognition domain of the regulator would provide valuable insight. The effector molecule of NdgR has not been identified despite several attempts (Kim et al. 2012b; Yang et al. 2009a). Instead, we explored a higher level of the NdgR regulatory network. Previous ChIP-chip experiments revealed *ndgR* as one of the targets of the oxidative stress response sigma factor SigR (Kim et al. 2012a). We measured mRNA expression level to validate the transcription of *ndgR* by SigR in response to thiol oxidative stress. NdgR was transiently induced by diamide treatment in the presence of SigR (**Figure 4.8A**); hence, NdgR is expected to regulate its target genes in response to thiol oxidative stress. However, NdgR was expressed constitutively in the absence of SigR, suggesting that there are additional transcriptional regulators in addition to the SigR.

Next, we sought to examine whether the *ndgR* mutant strain is sensitive to this thiol-reactive compound using a plate assay. We used complex media for the sensitivity test because *ndgR* mutant hardly grows in minimal media (Yang et al. 2009a; Kim et al. 2012b). When WT, *ndgR* mutant (BG11) and complemented mutant (BG13) spores were spotted on the R5- agar plates containing 0.6 mM diamide, BG11 cells were found to be sensitive to diamide even in the complex media (**Figure 4.8B**). These results show that *ndgR* is necessary for the response to oxidative stress conditions.

The role of BCAAs in response to stress conditions has not been completely elucidated in bacteria; however, *leuCD* induction in response to thiol-specific

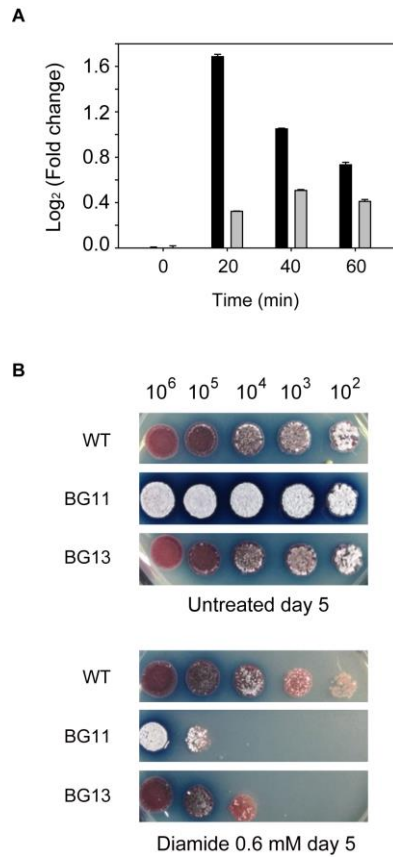


Figure 4.8 SigR-dependent transcription activation of the *ndgR* gene in *S. coelicolor*. (A) Quantitation of *ndgR* transcripts using quantitative realtime-PCR analysis (qRT-PCR) from diamide-treated cells reveals that the *ndgR* gene is induced using the SigR promoter. mRNA level of *ndgR* in WT (black bars) was induced under diamide treatment compared to the level of *ndgR* in *sigRrsrA* deletion mutant (gray bars). All levels are normalized by the levels of each sample at 0 min. (B) Sensitivity test of WT, *ndgR* deletion mutant (BG11) and complemented mutant (BG13) under thiol oxidative stress. Serially diluted spores of WT, BG11 and BG13 were spotted on R5 agar plates with or without added diamide (0.6 mM). Plates were incubated at 30°C for 5 days.

oxidative stress was reported in *Mycobacterium bovis* BCG (Dosanjh et al. 2005). In plants, there have been many reports regarding the accumulation of BCAAs in abiotic stress conditions (Obata and Fernie 2012; Araújo et al. 2010; Joshi et al. 2010). It has been suggested that BCAAs function as compatible osmolytes since the level of BCAAs is elevated under drought stress in various plant tissues (Joshi et al. 2010). Another possible role for BCAAs under stress conditions that has been suggested is that they function as alternative electron donors for the mitochondrial electron transport chain via the electron transfer flavoprotein (ETF) complex to produce ATP (Araújo et al. 2010). Isovaleryl-CoA from the degradation of BCAAs provides electrons to the ETF complex via the action of isovaleryl-CoA dehydrogenase. In addition, we observed that the *ndgR* mutant exhibited defective membrane formation in minimal media (data not shown). This result is likely due to the fact that 70% of total fatty acids in the membrane are branched-chain fatty acids, which are synthesized from the precursors derived from BCAA degradation (De Rossi et al. 1995).

Sulfur-related reactions are well known as anti-oxidation reactions in actinomycetes (Lee et al. 2005; Kim et al. 2012a; Paget et al. 2001). Expression levels of *cysIHCDN*, *cysM*, and *cysA* are significantly increased by SigB under osmotic and oxidative stresses (Lee et al. 2005), leading to an increase in cysteine levels. As a component of mycothiol, a major thiol buffer found in many actinomycetes, cysteine would protect the cell against osmotic and oxidative stresses. Because oxidation-labile S-containing factors such as TPP and iron-sulfur clusters are involved in many physiological reactions, NdgR would contribute to replenishing these factors. Thus, we speculate that NdgR maintains the intracellular redox balance and the structural

integrity of the membrane in response to external thiol oxidative stress by orchestrating the genes in its regulon.

4.7 Elucidation of NdgR regulatory logic

The members of the IclR family of regulators have been demonstrated to be activators, repressors, and dual-role proteins in many cases (Antonio et al. 2006). However, to ascertain the physiological role of a regulator, a comprehensive understanding of its regulatory modes, including higher and lower levels of regulation, would be helpful. In order to elucidate how NdgR regulates target gene expression in response to oxidative stress, we quantified mRNA levels of the relevant genes using qRT-PCR (**Figure 4.9**). We selected sulfur assimilation into the cysteine biosynthesis pathway as a target due to its significance in the stress response (Lee et al. 2005). The selected genes were *cysI* in *cysIHCDN* operon, *cysA* and *cysM* in cysteine biosynthesis, and *ndgR*. First, we confirmed that the transcriptional level of *ndgR* is induced by diamide. All of the target genes were also induced by diamide regardless of the presence of *ndgR*. This result indicates that all of the examined genes respond to oxidative stress, and that other factors such as SigR could exist. Next, we observed two regulatory modes based on the measurement of expression levels affected by NdgR under diamide treatment. *cysA* and *cysI* were induced by NdgR regardless of diamide. However, NdgR repressed the expression of *cysM*; thus, the role of NdgR as a dual regulator was confirmed.

We further explored the physiological roles of these two regulatory modes by NdgR using network motif theory (Mangan and Alon 2003). The two regulatory modes of NdgR were coherent type-1 feed-forward loop (C1-FFL) with OR-gate,

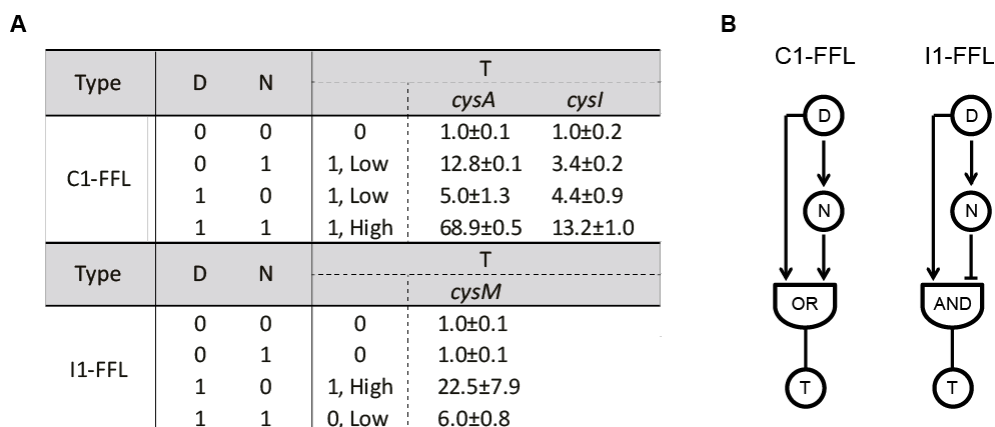


Figure 4.9 The regulatory modes of NdgR. (A) Measurement of expression levels of NdgR target genes in the sulfate assimilation pathway in various combinations of input signals. D, N and T denote diamide treatment, *ndgR* gene and target genes of NdgR, respectively. D = 0 or 1 indicates nontreatment or treatment of diamide, respectively. The absence or presence of the *ndgR* gene is denoted as 0 or 1, respectively. Expression of target genes above threshold are denoted as 0 (OFF) or 1 (ON) of output signals. Expression levels were normalized relative to the expression levels of controls (D=0, N=0). (B) The logic gates of NdgR regulatory networks. NdgR regulates the sulfate assimilation pathway using coherent and incoherent FFL.

and incoherent type-1 feed-forward loop (I1-FFL) (Fig. 5B). A C1-FFL is a regulatory pattern in which an activator (D) controls a target gene (T) and also activates another activator (N) of that target gene. C1-FFL with OR-gate shows a delayed response to OFF steps of D and a rapid response to ON steps (Mangan and Alon 2003). Because sulfur assimilation governed by *cysI* and *cysA* is controlled by C1-FFL with OR-gates, cells can maintain sulfur assimilation activity after the signal is OFF. Meanwhile, an I1-FFL is a regulatory pattern in which an activator (D) controls a target gene (T) and also activates a repressor (N) of that target gene. This motif is a pulse generator and a response accelerator (Mangan and Alon 2003). In addition, I1-FFL provides fold-change detection that responds only to the fold-change (rather than absolute change) of the input signal (Goentoro et al. 2009). CysM has an I1-FFL network motif; thus, it may be expressed in a rapid response that is proportional to the fold-change in the stimulus relative to the background.

We revealed that NdgR controls sulfur assimilation into the cysteine biosynthesis pathway through two regulatory modes. Using C1-FFL with OR-gate, NdgR can initiate a reduction in sulfate immediately in response to the stress and stably protect assimilation systems against transient loss of signal. At the final step, NdgR acts as a memory of stress intensity by using the I1-FFL motif. Thus, it mediates a continual temporal comparison between the present and past levels of stresses. During the stress condition, the memory is adjusted to the new level of stress intensity, and cysteine synthesis by CysM returns to its basal level. Thus, this modulation prevents excess synthesis of cysteine, which is energetically wasteful and avoids potential osmotic imbalances. Using these regulatory modes, NdgR likely provides an advantage for *S. coelicolor* to maintain homeostasis in stress conditions.

4.8 Conclusion

We demonstrated that this PCR-based tandem epitope tagging system is a versatile tool for investigating the regulatory roles of transcription factors in *Streptomyces coelicolor*. With a series of template plasmid which has tandem myc epitopes and apramycin resistance cassette, we can fuse the epitope tag into any desired chromosomal loci. Expression levels of target genes and phenotype showed that endogenous tagging did not change in vivo activity of the transcription factor of interest. In addition, the conservation of DNA binding activity of tagged protein was confirmed by ChIP-qPCR, so that this tool can be used for identifying direct binding sites at the genome-scale by combining with microarray or massively parallel sequencing, ChIP-chip or ChIP-seq, respectively (Gilchrist et al. 2009). Since the myc tag is a commonly used epitope, inexpensive antibodies are commercially available without making specific antibodies against the proteins of interest. Enhancement of detection sensitivity by the repeated epitope sequence allowed stringent conditions for washing to remove the nonspecific binding and to increase signal-to-noise ratio.

Next, we identified genome-wide binding sites of NdgR in *S. coelicolor* using ChIP-seq and revealed its physiological role under oxidative stress conditions. In our growth condition, we found that NdgR directly regulates 34 genes that are involved in the synthesis of BCAAs and cysteine using 19 regulatory binding sites. We confirmed that SigR, an oxidative stress response sigma factor, induces NdgR in response to thiol oxidative stress induced by diamide treatment. Interestingly, this implied physiological roles of the NdgR regulon in *S. coelicolor*. Degradation of

BCAAs is known to produce major CoA precursors of branched-chain fatty acids, which are the major components of bacterial cell membranes; thus, their production enhances the robustness of the cell. Furthermore, BCAAs can serve as an alternative electron transport donor for energy production under various stress conditions, similarly to that which occurs in plants. Because the induction of biosynthesis of BCAA under stress conditions has also been reported in other bacteria, the exact roles of BCAAs in stress conditions require further investigation. In addition, many BCAA biosynthesis genes require sulfur-containing cofactors such as thiamine and iron-sulfur clusters, which are vulnerable to thiol oxidative stress; these cofactors can be replenished by sulfur-related pathways such as cysteine biosynthesis, which is also regulated by NdgR. Furthermore, cysteine is one of the precursors of mycothiol, a major redox buffer in *S. coelicolor* that helps maintain redox balance in the cell. The cysteine biosynthesis pathway is regulated by NdgR under thiol oxidative stress using coherent and incoherent FFL, which enables cells to adapt to the environmental conditions by maintaining homeostasis. This study provides a deeper understanding the mechanisms by which NdgR regulates amino acid biosynthesis in response to stress in *S. coelicolor*. In addition, this model system shows a possibility such that revealing the global regulatory network of transcription factors using ChIP-seq technique will enable in depth understanding of their physiological roles in *S. coelicolor*.

Chapter 5. Transcriptional and translational landscape of *Streptomyces coelicolor* genome

5.1 Integration of genome-wide data generated by Next-generation sequencing technology

The high-resolution map of transcriptional and translational landscape of *S.coelicolor* genome was obtained through systematic integration of multiple NGS data (**Figure 5.1**). First step of the systematic integration was the production of genome-wide transcription start site (TSS) profiles. Intact RNAs carrying 5'-triphosphate group were distinguished from processed RNAs such as degraded mRNAs, mature rRNAs and tRNAs by using primary transcriptome sequencing technique, termed TSS-seq. To identify the most of TSSs in the *S. coelicolor* genome as much as possible, we obtained RNA samples from *S. coelicolor* cultures grown in 44 of different environmental conditions, including R5- complex liquid and solid media, carbon and nitrogen source combinations of minimal liquid media and various stress conditions (heat, cold, detergent, salt and ethanol). All extracted RNAs were pooled to determine TSSs at once. Combined RNA samples were divided for treatment and non-treatment of terminal exonuclease (TEX) to discriminate primary transcripts from processed transcript. Each experiment yielded sequencing reads, 1,472,916 and 3,978,037 with an average read length, 119.5 and 117.7 bp, respectively (**Table 5.1**). Total 3,926 TSSs, including TSSs of 3,468 protein-coding genes, were identified, which is the most comprehensive information about *S. coelicolor* TSS so far and agreed previously annotated TSSs (**Figure 5.2**).

We, then, measured global RNA expression level across the *S. coelicolor* genome

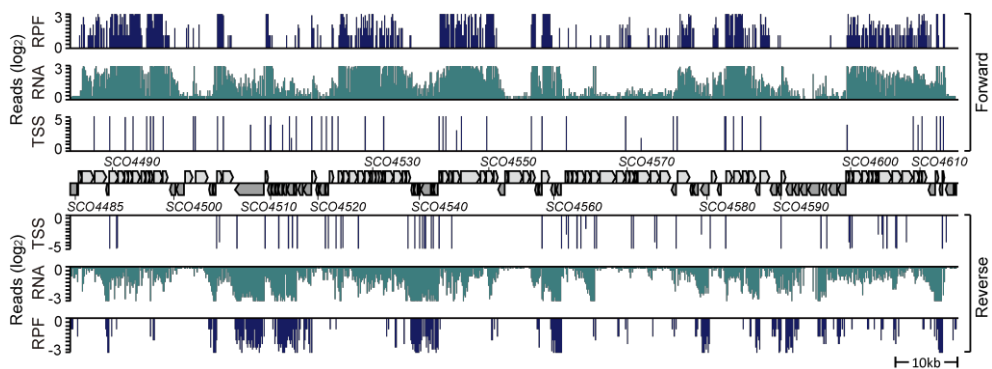


Figure 5.1 Integration of multiple high-throughput datasets mapped onto *S. coelicolor* genome. Data sets include transcription start sites (TSS), mRNA expression profiles (RNA), and RPF profiles (RPF). Ribosome profiling data was obtained by Minwoo Kim et. al.

Table 5.1 Sequencing statistics

Sample	RNA-seq								Ribo-seq				TSS-seq	
	EE1	EE2	ME1	ME2	LE1	LE2	S1	S2	EE	ME	LE	S	TEX+	TEX-
Total reads	10,143,803	17,337,483	16,538,119	16,557,033	14,994,962	15,350,264	17,292,840	15,849,845	16,426,951	17,926,930	19,970,666	19,399,084	1,472,916	3,978,037
Total mapped reads	8,207,090	14,747,001	14,615,494	14,693,793	12,672,436	12,339,943	15,468,256	12,777,492	15,985,272	17,680,226	19,600,724	18,945,548	1,028,872	3,559,540
Uniquely mapped reads	8,099,946	14,487,993	14,417,187	14,503,560	12,381,100	12,071,538	14,966,651	12,127,279	1,174,382	860,647	1,473,087	1,025,257	1,012,457	3,223,958
Reads mapped to CDS	7,155,940	12,703,910	12,636,228	12,675,489	10,727,085	10,227,791	12,306,722	10,125,896	808,741	491,389	851,450	704,685	670,859	2,469,276
Reads mapped to noncoding region	944,006	1,784,083	1,780,959	1,828,071	1,654,015	1,843,747	2,659,929	2,001,383	365,641	369,258	621,637	320,572	341,598	754,682
Read length in average (bp)	50.7	50.7	50.7	50.7	50.7	50.7	50.7	50.7	30.2	30.7	30.5	32.3	119.5	117.7

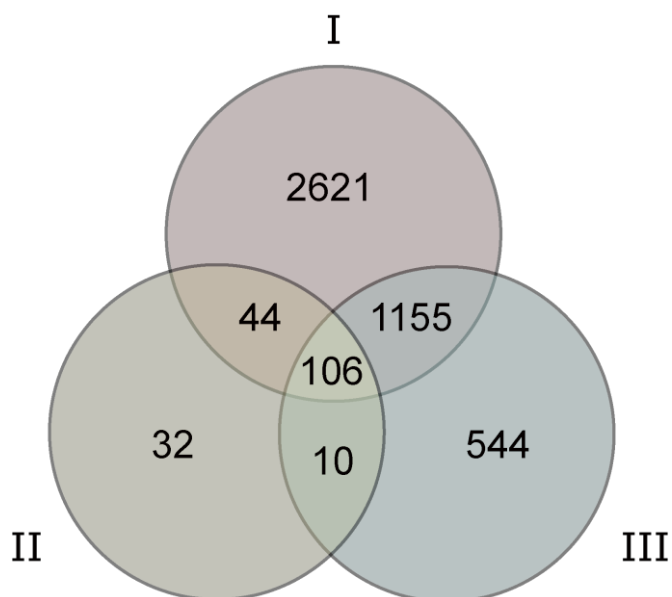


Figure 5.2 Comparison between previously known TSSs and the TSSs identified in this study. I, TSS identified in this study; II, TSS previously identified in the study of Vockenhuber et al. (Vockenhuber et al. 2011); III, TSS previously identified in the study of Romero et al (Romero et al. 2014).

using RNA-seq. RNA samples were extracted from *S. coelicolor* cultures grown in R5- medium in four different growth phases, early exponential (EE), mid-exponential (ME), late exponential (LE) and stationary (S) phase in duplicate (**Figure 5.3**). Massive-scale sequencing was performed to all eight samples and each output was greater than 10 million reads with an average read of 50 bp (**Table 5.1**). Sequence reads were aligned onto the *S. coelicolor* genome (NC_003888) to determine the number of reads matching each genomic position. In most samples, over 80% of reads were uniquely mapped to one genomic region. Mapping the reads to the genome allowed the determination of expression level of each gene at each time points. Many genes were differentially expressed depending on the growth phase. For example, gene cluster of red-pigmented antibiotics, prodiginine, was highly expressed from late exponential phase (**Figure 5.4**) as previously known (Huang et al. 2001).

The final step was integration of translome data with TSS and transcriptome data. The translome data was achieved from ribosome profiling technique, termed Ribo-seq, which is based on the deep sequencing of ribosome protected mRNA fragments (RPFs), performed by Minwoo Kim et. al (unpublished data). For the precise comparison with transcriptome data, the sequencing samples were obtained from the same cell cultures with strand-specific RNA-seq samples. The sequencing yielded over 16 million reads for each sample and 78, 68, 80, and 84% of protein-coding genes were aligned with at least one read at early exponential, mid-exponential, late exponential and stationary phase, respectively. However, uniquely mapped reads were under 10% of total sequencing reads because most of reads

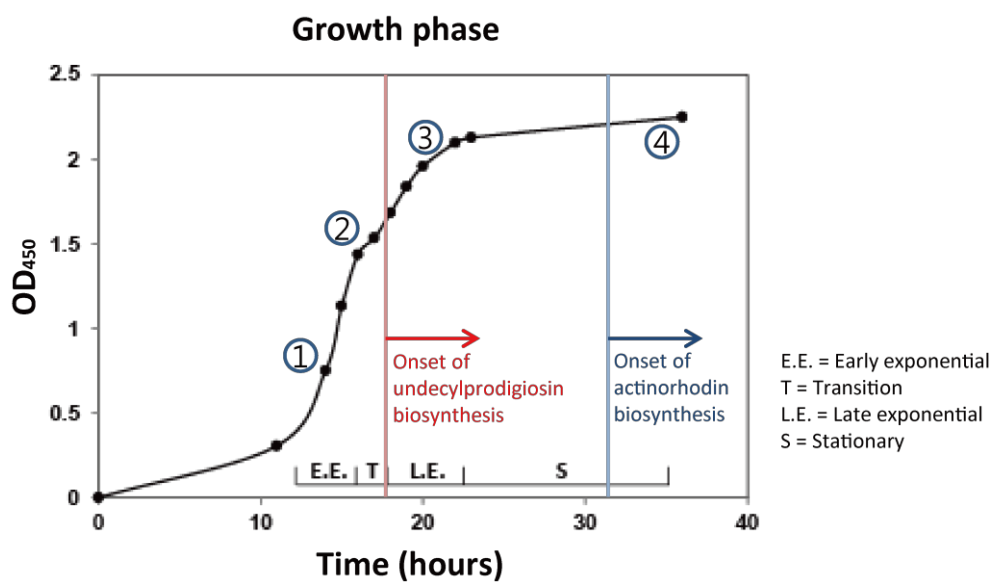


Figure 5.3 Growth phases of *S. coelicolor* and sampling point for this experiment.

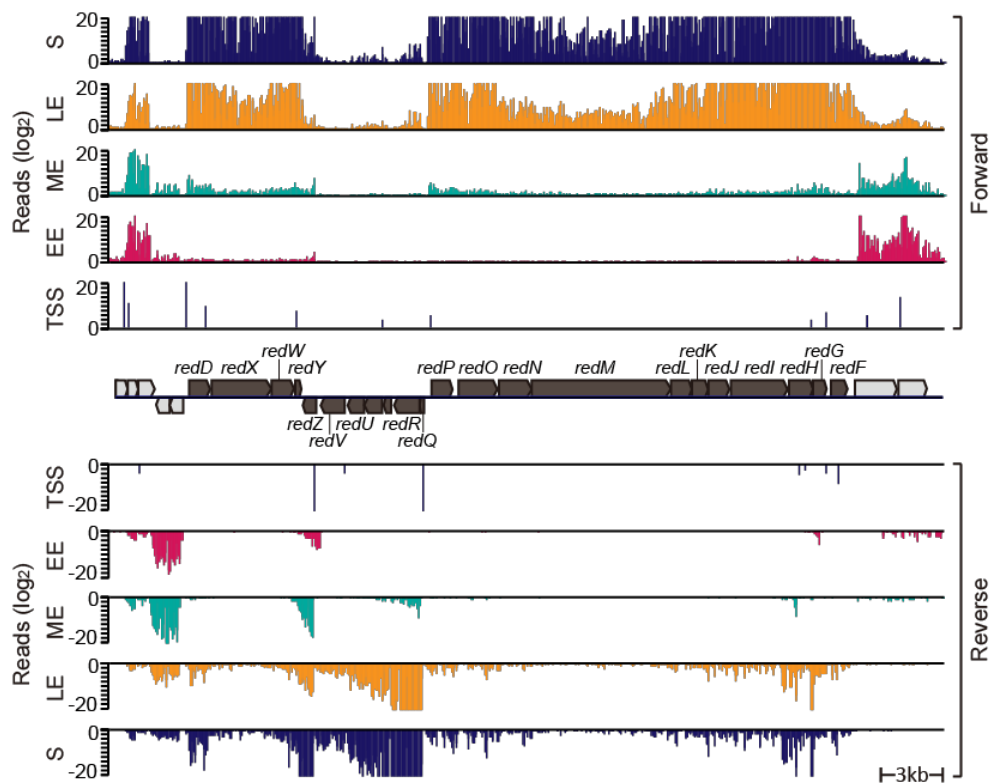


Figure 5.4 Differential RNA expression of prodiginine gene cluster (brown arrow).

EE, early exponential; ME, mid-exponential; LE, late exponential; S, stationary.

were matched onto rRNAs. Compared with mRNA expression, the correlation coefficient (R^2) was 0.73, 0.63, 0.73, and 0.79 at early exponential, mid-exponential, late exponential and stationary phases, respectively (**Figure 5.5**). The normalized occupancy of RPF on each gene was defined as protein expression level.

Integration of these three NGS data leads us to analyze the complex genome architecture of *S. coelicolor*, such as discovering novel sRNAs or ORFs in secondary metabolite gene clusters (**Figure 5.6**). Furthermore, comparing transcriptomic and translational data provides the dynamic expression level of *S. coelicolor* according to its developmental stages.

5.2 High-resolution map of genetic organizational elements

Massive and single-based resolution of sequencing data was analyzed to decipher various structural elements of *S. coelicolor* genome.

First, 3926 TSSs identified from TSS-seq were categorized by their positions (**Figure 5.7**). TSSs between 500 bp upstream and 150 bp downstream of a start codon of open reading frames (ORFs) were designated as primary TSSs (P) and TSSs located inside of ORFs were designated as internal TSSs (I). Meanwhile, TSSs on the opposite strand of ORFs were designated as antisense TSSs (A), and the others that are not associated with any other categories described above were designated as orphan TSSs (O).

Among the categorized TSSs, the length of 5' untranslated region (UTR) of each gene could be calculated from 3468 primary TSSs (**Figure 5.8**). The genes having 0 to 9 nt length of 5' UTR were determined as leaderless genes, about 20% of

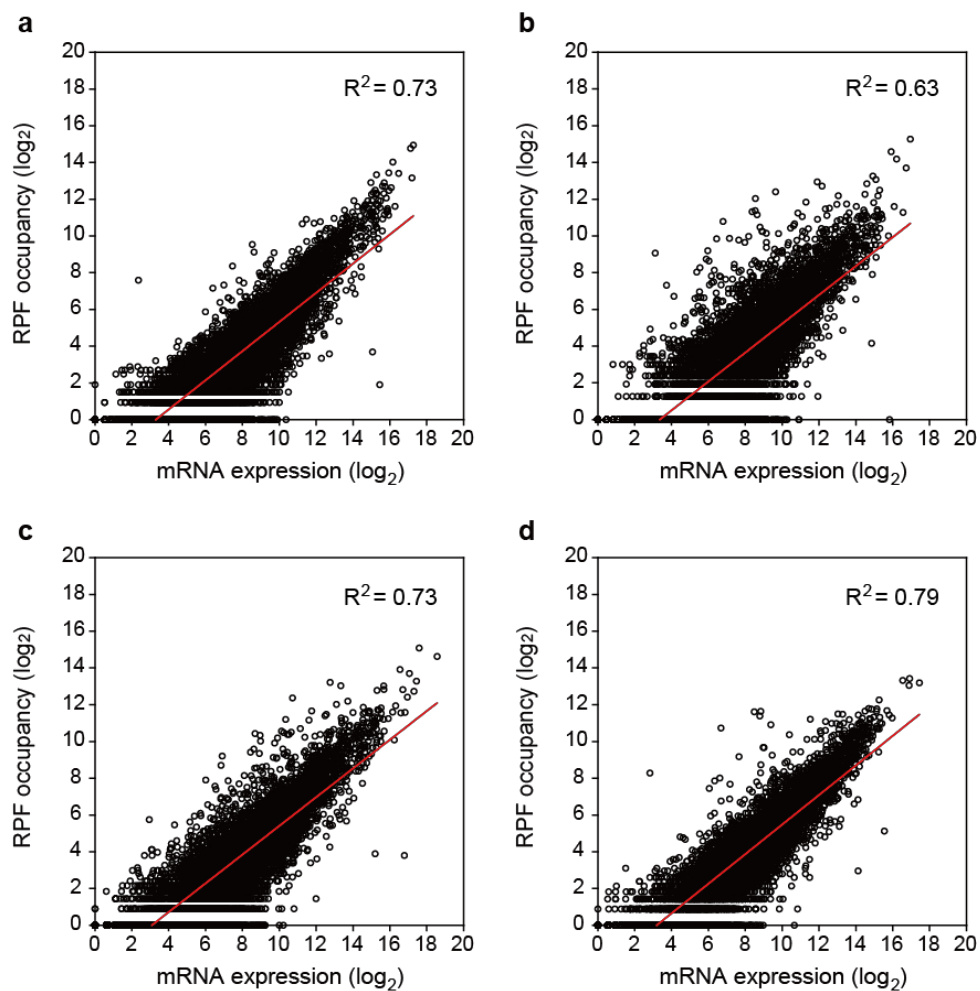


Figure 5.5 Correlation between RNA-seq and Ribo-seq data. (a) Early exponential phase (EE) (b) Mid-exponential phase (ME) (c) Late exponential phase (LE) (d) Stationary phase (S)

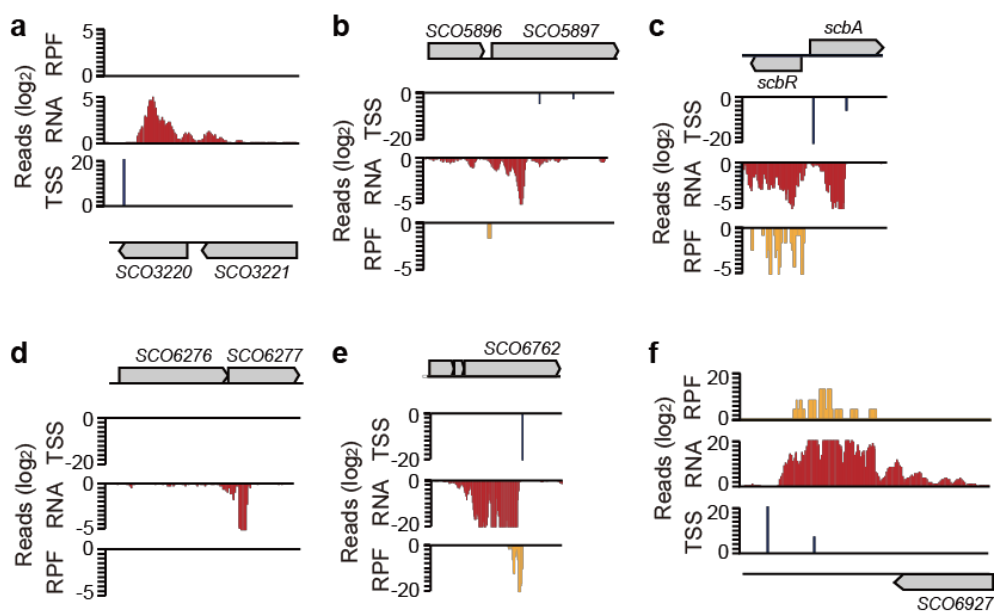


Figure 5.6 Novel sRNAs and ORFs within secondary metabolite gene clusters. Antisense RNAs within (a) CDA cluster (SCO3210-3249), (b) prodiginine cluster (SCO5877-5898), (c) SCB1 (SCO6266), (d) coelimycin P1 cluster (SCO6273-6288), (e) hopene cluster (SCO6759-6771) and (f) lantibiotic cluster (SCO6927-6932).

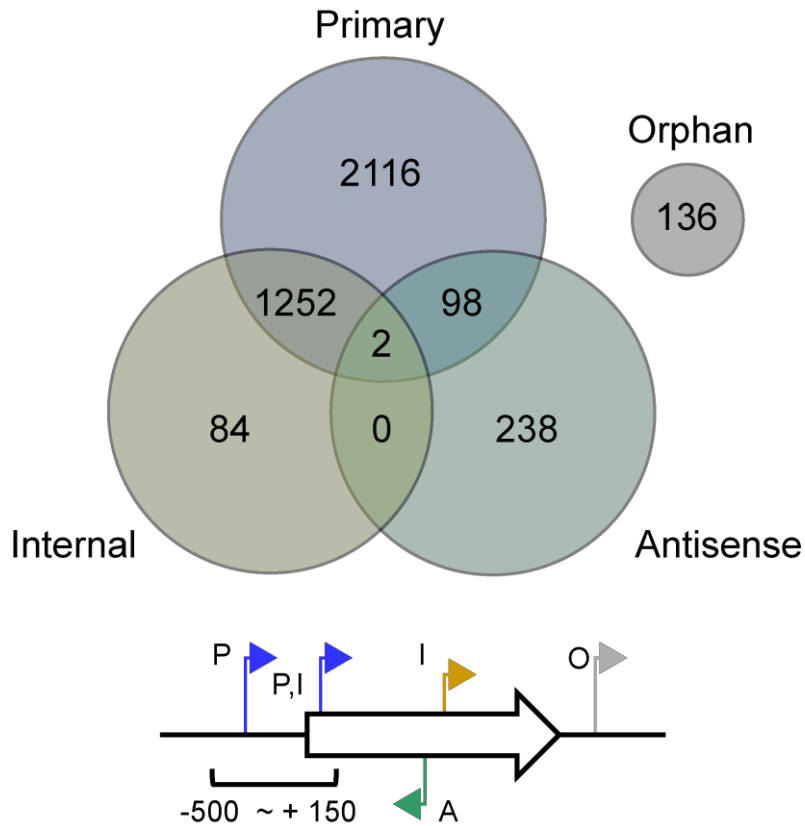


Figure 5.7 Identified TSSs classified by their positions. TSSs located from 500 bp upstream to 150 bp downstream of annotated start codon of ORF was classified to primary (P). In the case of the TSSs located within annotated start to stop codon, the TSS on the same strand with the ORF was classified to internal (I), and on the opposite strand was antisense (A). The TSSs not included any of the categories mentioned above were classified to orphan (O).

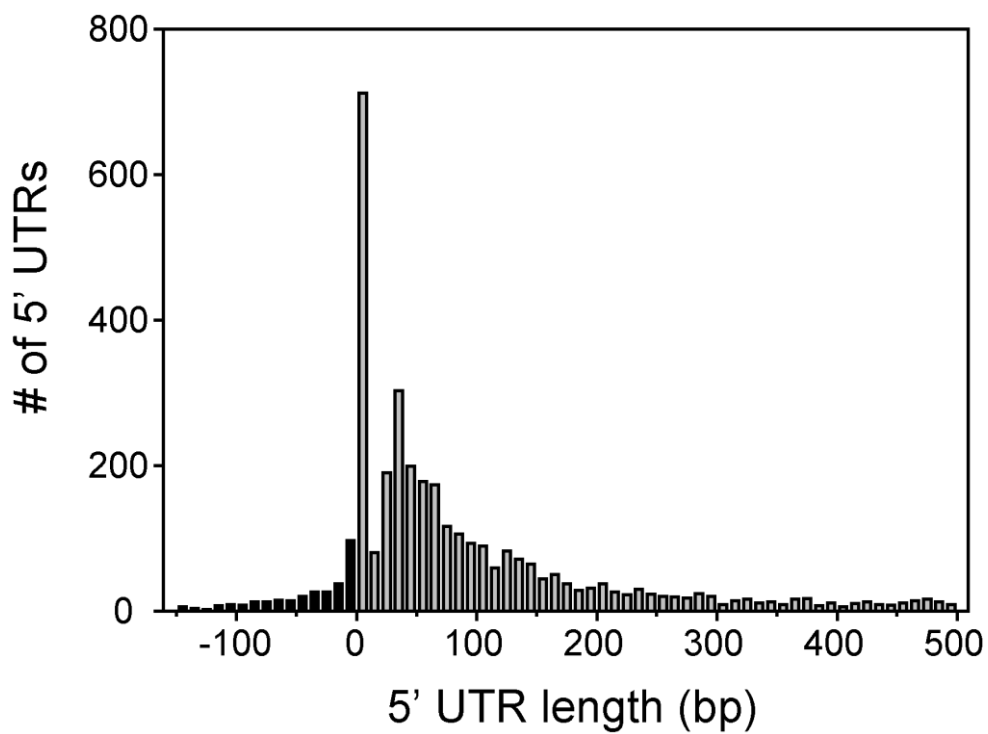


Figure 5.8 Distribution of 5' UTR lengths. Genes with the 5' UTR length of 0 to 9 nucleotides, considered as leaderless, were the most frequent.

primary TSSs, and the most frequent interval of UTR length except leaderless genes was 30 to 39 nt, This result agreed to previous report that predicted proportion of leaderless gene in *S. coelicolor* (Zheng et al. 2011). Relatively abundant existence of leaderless genes in actinomycetes, antibiotics producers, was predicted (**Appendix I**) because many antibiotics inhibit translation initiation on leadered transcripts.

To examine the functional features of leaderless genes, functional categories of leaderless genes were analyzed using Clusters of Orthologous Groups (COGs) (**Figure 5.9**). By COG analysis, a noticeable difference of the ratio was revealed in transcription category; 25 % of the leaderless genes were characterized to the transcription category while 14 % of total *S. coelicolor* genes were related to transcription. The reason why transcription category has abundant in leaderless genes need to be studied.

We compared translational efficiency (TE: RNA level over RPF abundance) changes to the length of 5' UTR (**Figure 5.10**). TE was increased according to the growth phase in leaderless genes while TE value of genes with 5' UTR had no change in TE level regardless of the length and growth phases. With the presence of 5' UTR, the change of TE by several factors might be canceled out while leaderless genes were controlled by common initiation factors similar to previous report that translations of leaderless genes at stationary phase in *Mycobacterium tuberculosis* were stimulated (Cortes et al. 2013).

TSSs were then used to find a promoter motif. Between 30 nt upstream and 15 nt downstream sequences from TSSs were collected and investigated (**Figure 5.11**). Conserved sequences were detected at about -7 (T) and -11 (A) position from TSSs,

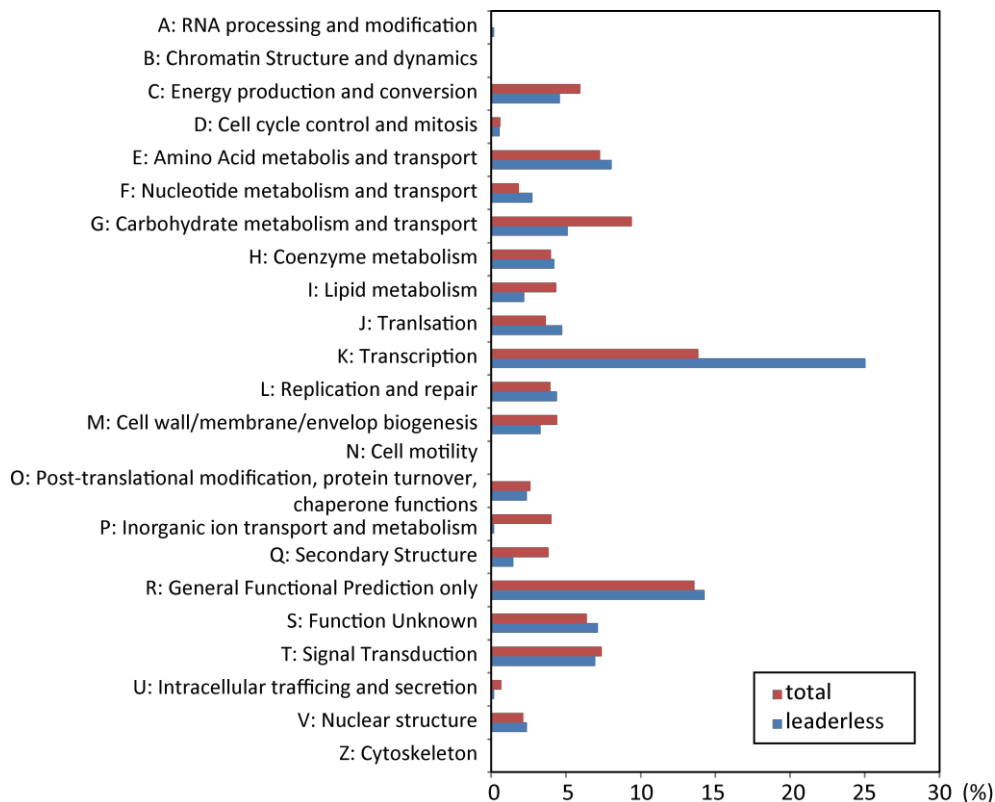


Figure 5.9 COG functional categories of leaderless genes. Each bar represents the ratio of the number of total genes and leaderless genes involved in corresponding category compared to the number of total genes (red) and total leaderless genes (blue), respectively.

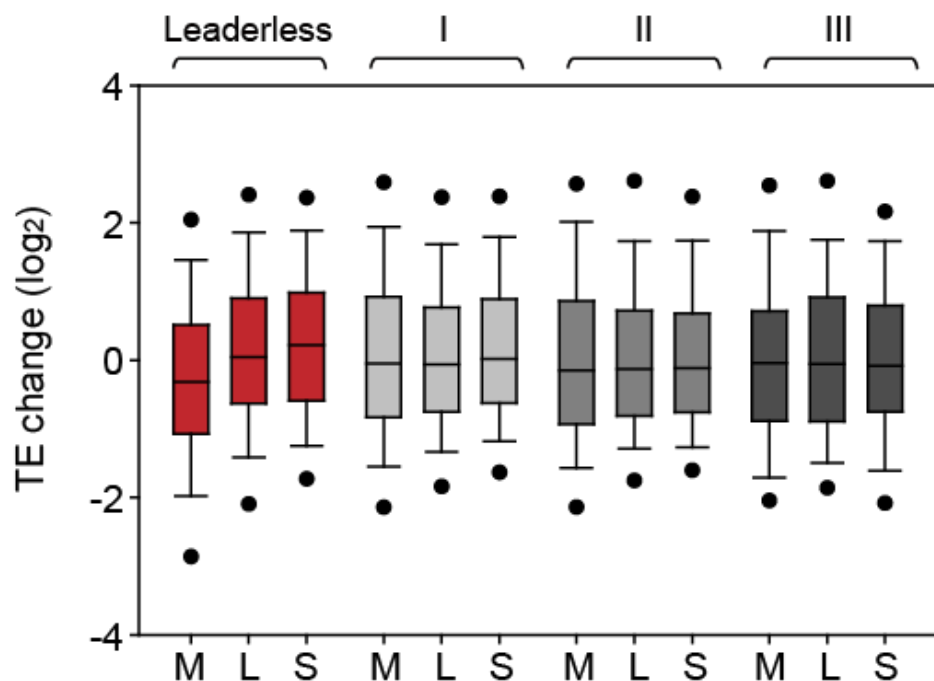


Figure 5.10 TE change distributions according to 5' UTR length and time points. Leaderless, 0 to 9 nt 5' UTR; I, 10 to 100 nt 5' UTR; II, 101 to 200 nt 5' UTR; III, 201 to 500 nt 5' UTR. M, fold change between early and mid-exponential phases; L, fold change between early and late exponential phases; S, fold change between early exponential and stationary phases.

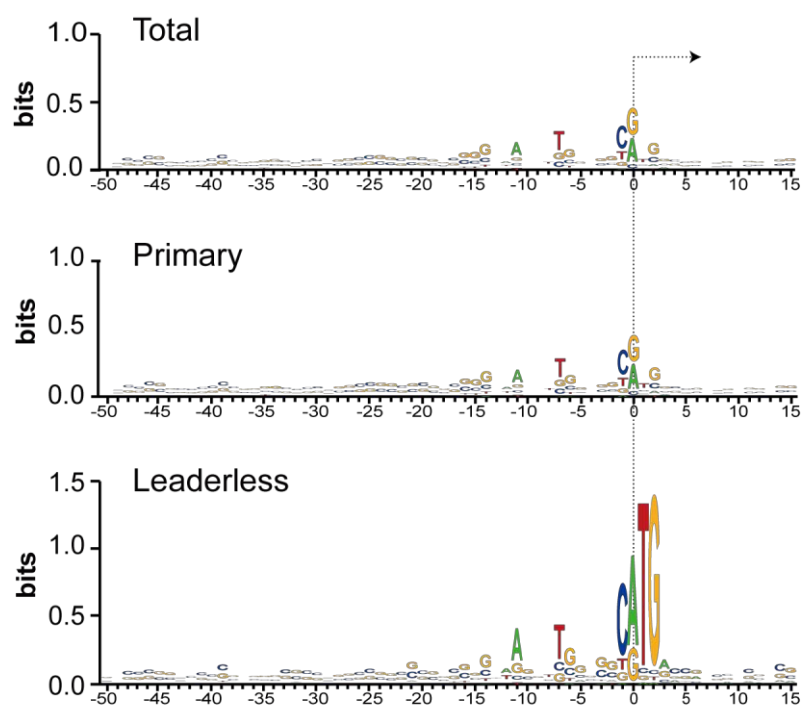


Figure 5.11 Motif sequences drawn with total TSSs or the specific subset of total TSSs; primary TSSs and leaderless TSSs identified in this study. The results show conserved motif sequences around -10 region in all data sets.

and the sequence is consistent with each subsets such as total TSS, primary TSSs and leaderless genes.

Meanwhile, 151 novel transcripts were discovered by analyzing RNA-seq profiles, and 79 of them were also found in previous report (Moody et al. 2013). We categorized them as antisense or intergenic transcripts according to their genomic positions. Furthermore, by integrating Ribo-seq profiles, we revealed 42 novel transcripts occupied by ribosomes, which were regarded as putative new ORFs, which are expected to have biological roles in *S. coelicolor*.

5.3 Discrepancy in mRNA and protein expression

To investigate growth phase responsive gene expression relationship between mRNAs and proteins in *S. coelicolor*, cluster analysis was performed using abundance of mRNA and RPF with Pearson correlation. The genes were divided into 11 groups according to expression pattern (**Figure 5.12**). In mRNA data, the groups of genes with increasing level according to the growth phase were group 1, 3, and 6 while the group of genes with decreasing level was group 7. In RPF abundance, genes with increasing level were involved in group 1 and 2 while genes with decreasing level were involved in group 10. To examine the groups having similar patterns of mRNA and RPF abundances, we designated as II for mRNA-increasing and RPF-increasing groups, and DD for mRNA-decreasing and RPF-decreasing groups (**Figure 5.13**). In group II, many genes in secondary metabolism are included as expected (**Figure 5.14**). In addition, many genes related to regulation such as RNA polymerase core enzyme binding and other transcription factors were also increased. The major categories in group DD were nucleotide

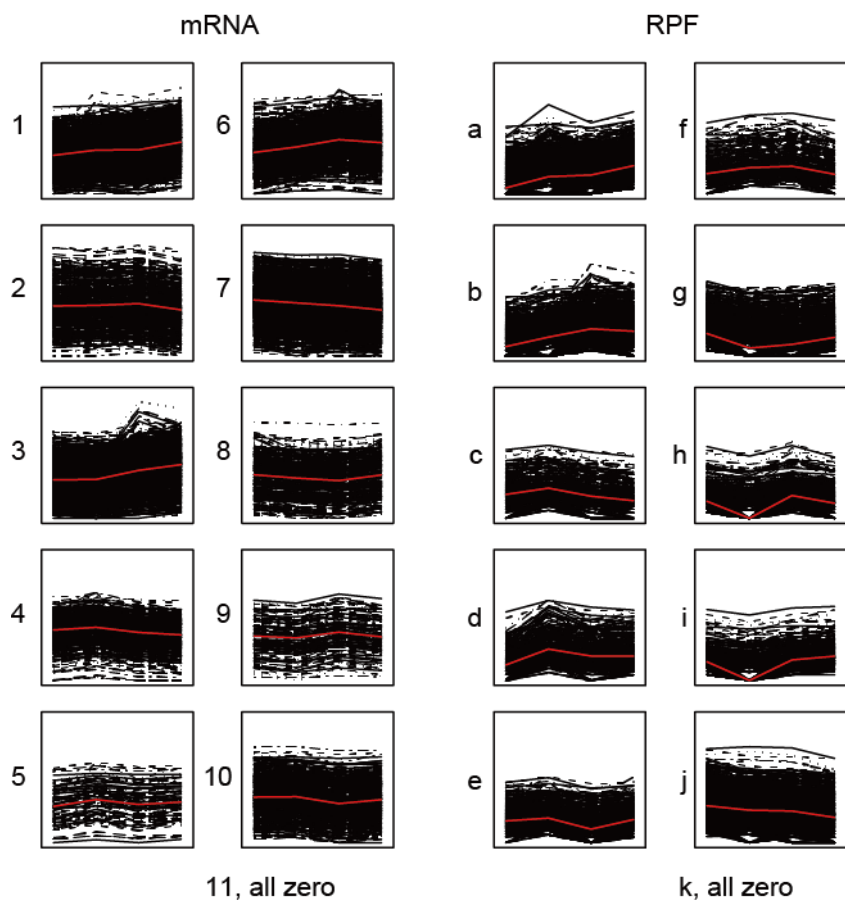


Figure 5.12 Cluster analysis of abundance of mRNA and RPF by Pearson correlation. Red lines indicate median values. X-axis is time point and Y-axis is normalized expression level.

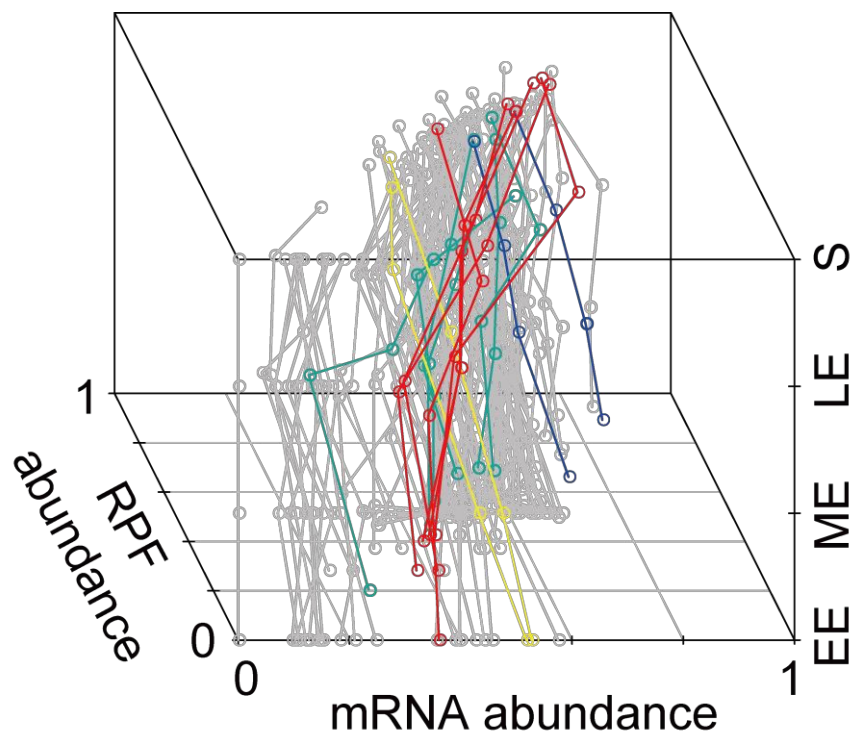


Figure 5.13 Expression pattern according to the median values of each cluster. Red lines indicate II groups, blue lines indicate DD groups, cyan lines indicate ID groups and yellow lines indicate DI groups.

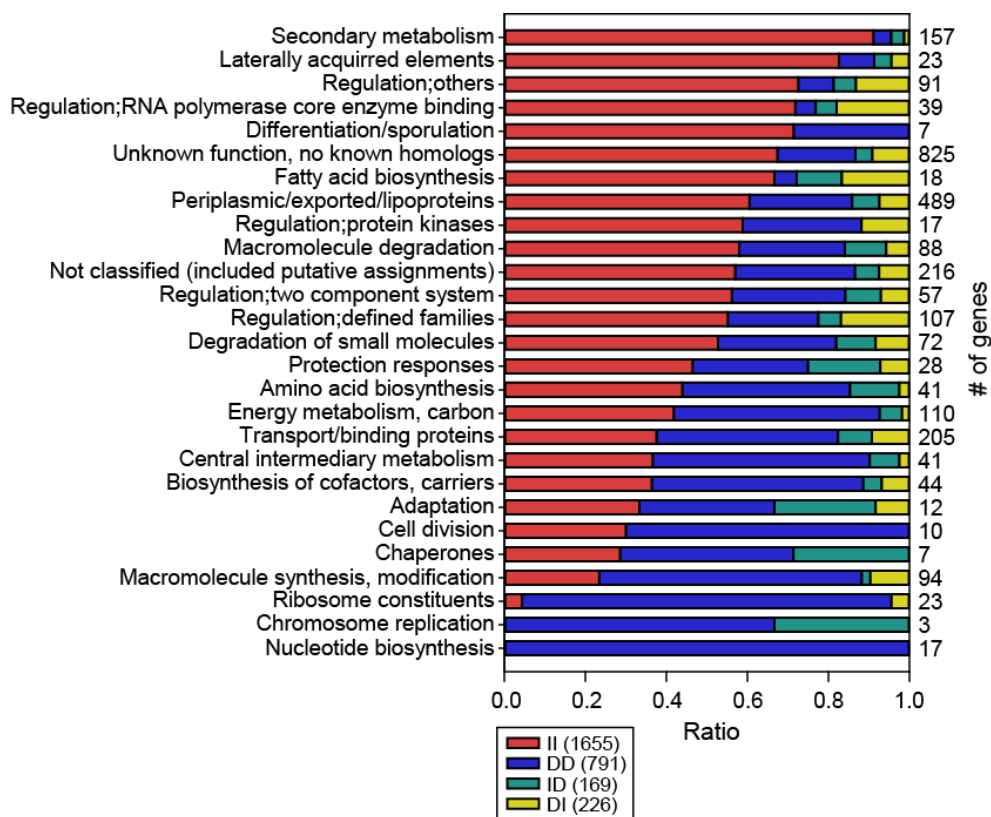


Figure 5.14 Functional enrichment analyses of the group of genes according to expression trend. Genes in each group (II, DD, ID and DI) were categorized into functional groups using functional classification from Sanger Institute database and plotted according to the ratio.

synthesis, ribosome constituents, cell division, and chromosome replication and so on, which are involved in cell growth or viability.

The significant number of genes showed anti-correlation between mRNA and RPF level. In group ID (mRNA-increasing and RPF-decreasing group), genes involved in chromosome replication and chaperones possessed major proportions (**Figure 5.14**). Meanwhile, some genes showed very exceptional trend, that is, group DI, which have decreasing trend in mRNA abundance and increasing trend in RPF abundance. Many genes involved in these groups have regulatory functions, which are RNA polymerase core enzyme binding, regulation of defined families or others, protein kinases and so on. Especially, many anti-sigma factors in RNA polymerase core enzyme binding were included. Although further studied are needed, these gene were likely controlled at translational level.

5.4 Genomic landscape of secondary metabolite genes in *S. coelicolor*

Integration of multiple NGS data analyzed above enabled us to describe the transcriptional and translational landscape of secondary metabolite genes in *S. coelicolor*, including structural and qualitative information.

First, we discovered 80 TSSs, including 72 primary TSSs, in 21 secondary metabolite gene clusters (**Figure 5.15, Appendix II**). The accurate identification of TSS makes it possible to examine transcriptional unit architecture, which enables us to manipulate genes more precisely for engineering secondary metabolite production. Next, to confirm the expression pattern of secondary metabolite genes, the distribution of expression levels of RNAs and proteins in total genes and secondary

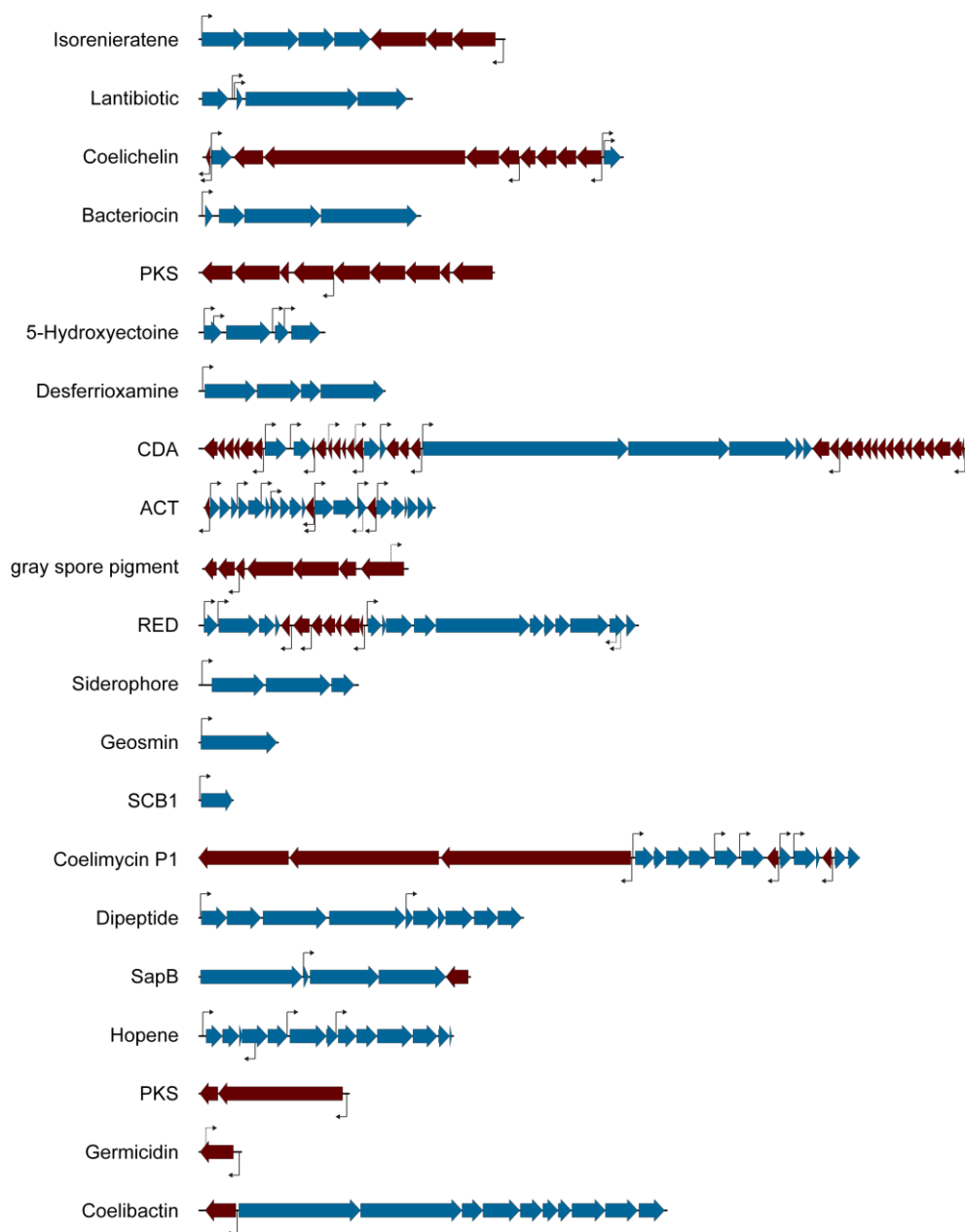


Figure 5.15 TSSs in secondary metabolite gene clusters identified in this study.

metabolite genes were compared. Expression changes between early exponential phase and each other time points were calculated, and distribution of the calculated values of each time points were represented separately (**Figure 5.16**). Median values of total gene expressions were nearly consistent at both mRNA and RPF levels regardless of the time points. This might be due to the level of up-regulated genes and down-regulated genes nullified each other, thus median values converged to zero. In contrast, median values of secondary metabolite gene expressions were increased gradually along the growth phases at both mRNA and RPF levels, and this trend was consistent with the clustering analysis.

For detailed investigation about expression of each secondary metabolite gene, we compiled whole expression profiles of secondary metabolite genes in *S. coelicolor*, which include 28 secondary metabolite gene clusters consisted of 221 genes (Craney et al. 2013) (**Figure 5.17, Appendix III**). mRNA abundances, RPF abundances, and TE changes of each gene were represented at each time points. Chemical structures of the products were denoted if they were known. This enabled us to examine the expression trend of each secondary metabolite gene cluster at once. For example, the gene cluster producing 5-hydroxyectoine, which plays an important role as stress protectants (Bursy et al. 2008), is highly expressed in both transcription and translation levels at every time points. Actual 5-hydroxyectoine production was confirmed with LC-MS (**Figure 5.18**). However, many of gene clusters not identified their structures were expressed at low levels at transcription and translation level. These kinds of genes such as three anonymous PKS gene clusters (SCO1265-1273, SCO6826-6827, and SCO7669-7671) were seldom expressed both in transcription and translation levels during the experimental

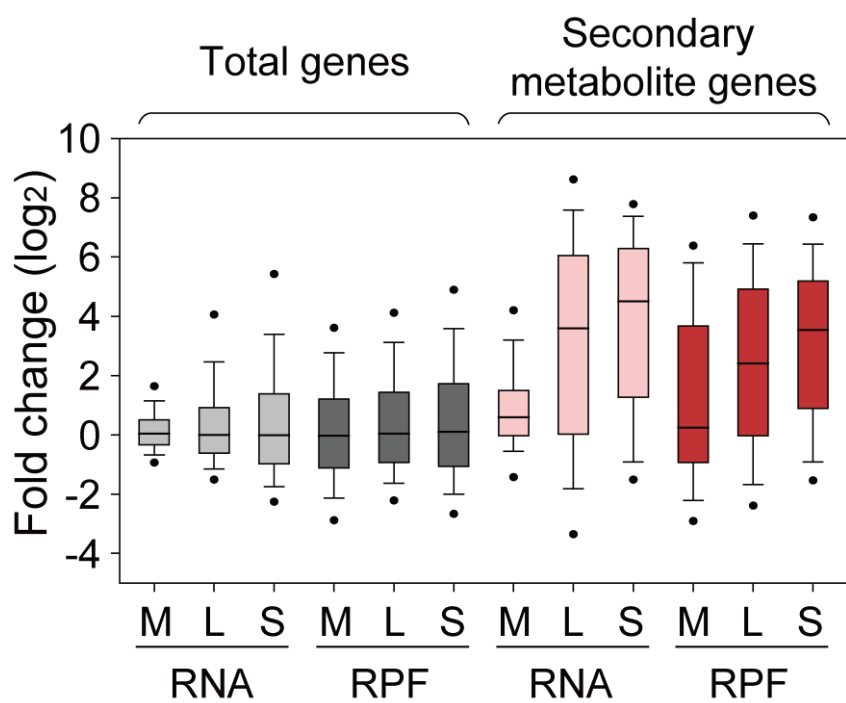


Figure 5.16 Distribution of mRNA fold change (mRNA) and RPF fold change (RPF).

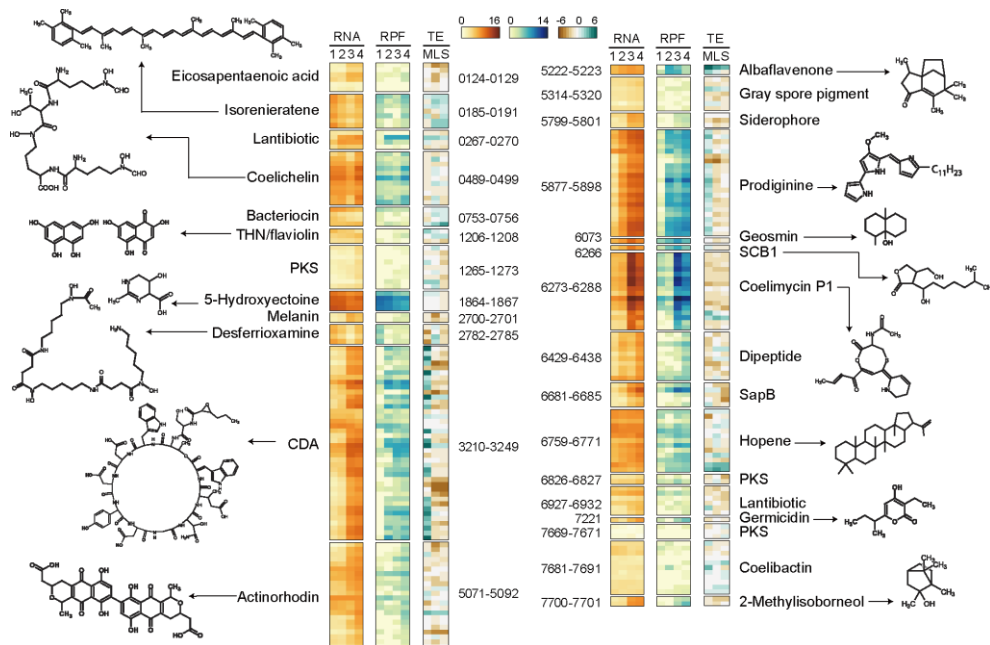


Figure 5.17 Heatmap of whole secondary metabolite genes in *S. coelicolor*. mRNA expression levels (mRNA), RPF occupancies (RPF) and the fold changes of translation efficiencies (TE) of 221 genes for 28 secondary metabolites were transformed to log2 scale. Outmost numbers are SCO numbers of the genes included in each clusters. Chemical structure was indicated for known secondary metabolite structures. 1, early exponential phase; 2, mid-exponential phase; 3, late exponential phase; 4, stationary phase; M, fold change between early and mid-exponential phases; L, fold change between early and late exponential phases; S, fold change between early exponential and stationary phases.

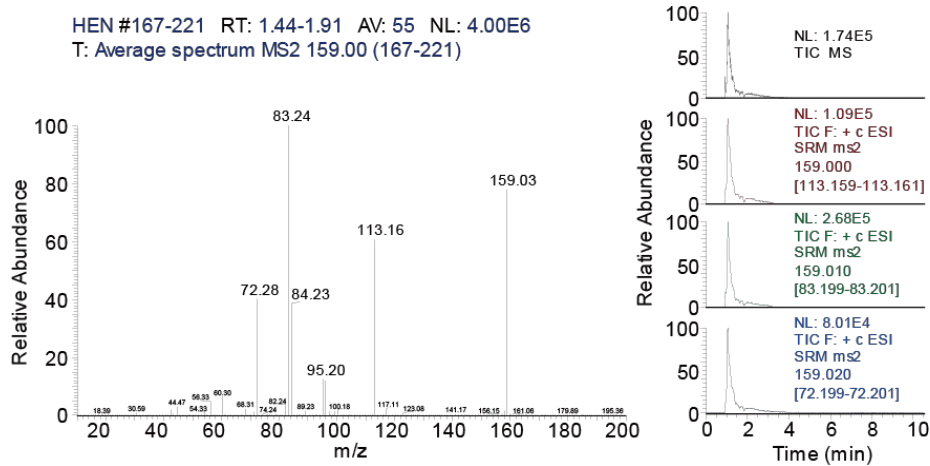


Figure 5.18 LC-MS data of 5-hydroxyectoine.

conditions. These metabolites have not been identified yet in spite of indomitable effort of secondary metabolite discoveries. This may indicate that these clusters are involved in cryptic or latent pathways in general laboratory experimental conditions. Meanwhile four representative antibiotics of *S. coelicolor*, calcium dependent antibiotics (CDA), actinorhodin, prodiginine and coelimycin, showed dynamic expression patterns depending on the growth phases. The expression levels of actinorhodin and prodiginine gene clusters were generally increased along the time points, while coelimycin gene clusters represented maximum expressions at late exponential phase. In case of CDA gene clusters, although mRNA expression levels were increased from late exponential phase, RPF abundance seemed to be increased earlier, at mid-exponential phase. Consequently, the expression changes at early to mid-exponential phase showed high TE values in CDA gene clusters. Furthermore, we were able to confirm that RedZ (SCO5881) and CdaR (SCO3217), a pathway specific regulator of prodiginine and CDA, respectively, were highly expressed prior to increase of the expressions of other genes in the same clusters, as previously reported (Huang et al. 2001).

With these comprehensive analyses, we could identify the genome architecture and expression patterns of secondary metabolite genes in *S. coelicolor* at systems level along the growth phases. Specifically, compared to transcriptional expressions, the moderate increase of translational expressions in secondary metabolite genes was observed, suggesting that synthesis of desired protein or metabolite can be improved through manipulating post-transcriptional regulation.

5.5 Conclusion

In this chapter, structural organizations and expression profiles of genes in *S. coelicolor* were investigated at transcriptional and translational level using multiple genome scale analysis. We integrated genome-wide data of transcription start sites (TSS-seq), mRNA abundance (RNA-seq), and ribosome-protected mRNA fragment (RPF) abundance (Ribo-seq), providing plentiful information about various structural constituent and expression control in *S. coelicolor* genome. By identifying 3926 TSSs in the genome, the features of 5' UTR of genes was deduced, which revealed abundant existence of leaderless genes (~20%). The major proportion of the function of leaderless genes is transcription and their TE is gradually increased along the growth phase. This may shows the possibility of undiscovered transcriptional regulation in stationary phase through leaderless genes in *S. coelicolor*. In addition, a number of putative noncoding RNAs (ncRNAs) and new ORFs being translated were detected.

In particular, disparity between the expression of mRNA and protein was examined derived from translational efficiency (TE), the level of RPF abundance over mRNA abundance. Although transcriptional regulation has been studied as one of the primary means of gene expression regulation, translational control, represented by TE, was revealed more complex than the previous thought. Especially, disagreement between transcription and translation rates in secondary metabolite genes was revealed. Understanding transcriptional and translational regulation and TSS profiles of secondary metabolite genes will be essential information to engineering of genetic circuits for the antibiotics synthesis in *S. coelicolor*.

Chapter 6. Conclusion & Further Suggestions

The amount of genomic information of bacteria has been rapidly increased with the development of sequencing technology. The genomic basis of secondary metabolite biosynthesis of *Streptomyces* has also been uncovered by genome sequencing. In this thesis, the various functional elements of *Streptomyces* genome was examined using genome-wide analyses based on the next-generation sequencing technologies such as ChIP-seq, TSS-seq, RNA-seq and Ribosome profiling.

First, the features of *Streptomyces* genome were investigated using pan-genome analysis. Comparative analysis of 17 completed *Streptomyces* genome provided core genome of this genus, especially sigma factors, cell division proteins and secondary metabolites biosynthesis genes that are distinct features of *Streptomyces*.

Next, tandem epitope tagging system was developed for elucidating transcriptional regulation using ChIP experiments. This can reduce time and cost of experiments, therefore it can be applied to the studies of various transcriptional regulators. In this thesis, the tagging was employed to identify genome-wide binding region of NdgR, a common amino acid biosynthesis regulator of *Streptomyces*. This revealed that NdgR controls branched-chain amino acids and sulfur-containing amino acids. Especially it was proved that NdgR maintains homeostasis of sulfur assimilation pathway using feed-forward loop by interaction with SigR, an oxidative stress response sigma factor, under thiol oxidative stress conditions.

Finally, to reveal the structural organizations and expression profiles of genes in *S. coelicolor*, multiple genome-wide data including transcription start sites (TSS-seq), mRNA abundance (RNA-seq), and ribosome-protected mRNA fragment (RPF) abundance (Ribo-seq), were integrated. Total 3926 of TSS were identified and the determination of 5' UTR length revealed a number of leaderless genes (~20% out of

primary TSSs). Many leaderless genes were involved in transcription category and the translational efficiency was increased reaching to the stationary phase. And the conserved promoter motif was able to be deduced from the identified TSSs. Dynamic change of RNA and RPF level showed disparity between transcription and translation, indicating the existence of translational control. In particular, transcriptional and translational landscapes of secondary metabolite genes were investigated deeply. With the integration of multiple NGS data, the single-based resolution map of genome architecture and expression profiles of each secondary metabolite clusters were examined, which provides valuable information for manipulating secondary metabolite production.

Taken together, the structure and dynamic expression changes of *S. coelicolor* genome were investigated at the genome-wide levels using multiple NGS tools. Moreover, regulatory network of transcription factor and translational control were discovered deeply. This information can be applied to developing strategies for antibiotics production, genetic manipulation and gene regulation. Genome-wide binding map of various DNA-binding proteins using epitope tagging system developed in this thesis would enlarge the transcriptional regulatory network of *Streptomyces*. Identification of TSS would make it possible to manipulate *Streptomyces* genes more precisely, which can be used for synthetic biology approach of this genus. Furthermore, accurate measurements of RNA and protein level in genome-scale would help to establish gene targets to be controlled by engineering promoters or RBSs. Information from integrated omics data and the methodologies of this thesis can be applied to other bacteria for system-level understanding of genome elements.

References

- Adessi C, Matton G, Ayala G, Turcatti G, Mermod JJ, Mayer P, Kawashima E (2000) Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Res* 28 (20):E87
- Al-Bassam MM, Bibb MJ, Bush MJ, Chandra G (2014) Response Regulator Heterodimer Formation Controls a Key Stage in *Streptomyces* Development. *PLoS genetics*. doi:10.1371/journal.pgen.1004554
- Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome biology* 11:R106
- Antonio JM-H, Tino K, Maria Eugenia G, Ana S, Juan LR (2006) Members of the IclR family of bacterial transcriptional regulators function as activators and/or repressors. *FEMS Microbiology Reviews*. doi:10.1111/j.1574-6976.2005.00008.x
- Araújo WL, Ishizaki K, Nunes-Nesi A, Larson TR, Tohge T, Krahnert I, Witt S, Obata T, Schauer N, Graham IA, Leaver CJ, Fernie AR (2010) Identification of the 2-hydroxyglutarate and isovaleryl-CoA dehydrogenases as alternative electron donors linking lysine catabolism to the electron transport chain of *Arabidopsis* mitochondria. *The Plant cell* 22 (5):1549-1563. doi:10.1105/tpc.110.075630
- Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic acids research* 37 (Web Server issue):8.
- Balleza E, López-Bojorquez LN (2009) Regulation by transcription factors in

- bacteria: beyond description. FEMS microbiology review 33(1):113-51
- Baranasic D, Gacesa R, Starcevic A, Zucko J, Blazic M, Horvat M, Gjuracic K, Fujs S, Hranueli D, Kosec G, Cullum J, Petkovic H (2013) Draft Genome Sequence of *Streptomyces rapamycinicus* Strain NRRL 5491, the Producer of the Immunosuppressant Rapamycin. Genome Announcements 1(4):e00581-13
- Barrell BG, McCormick JR, Santamaria RI (2004) SCP1, a 356 023 bp linear plasmid adapted to the ecology and developmental biology of its host, *Streptomyces coelicolor* A3 (2). Molecular microbiology 51(6):1615-28
- Bentley SD, Chater KF, Cerdeño-Tárraga AMM, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H, Harper D, Bateman A, Brown S, Chandra G, Chen CW, Collins M, Cronin A, Fraser A, Goble A, Hidalgo J, Hornsby T, Howarth S, Huang CHH, Kieser T, Larke L, Murphy L, Oliver K, O'Neil S, Rabinowitsch E, Rajandream MAA, Rutherford K, Rutter S, Seeger K, Saunders D, Sharp S, Squares R, Squares S, Taylor K, Warren T, Wietzorrek A, Woodward J, Barrell BG, Parkhill J, Hopwood DA (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature 417 (6885):141-147. doi:10.1038/417141a
- Bibb MJ (2005) Regulation of secondary metabolism in streptomycetes. Current opinion in microbiology 8 (2):208-215. doi:10.1016/j.mib.2005.02.016
- Bottacini F, O'Connell Motherway M, Kuczynski J, O'Connell KJ, Serafini F, Duranti S, Milani C, Turrone F, Lugli GA, Zomer A, Zhurina D, Riedel C, Ventura M, van Sinderen D (2014) Comparative genomics of the *Bifidobacterium breve* taxon. BMC genomics 15:170. doi:10.1186/1471-

- Brolle D, Bentley S, Kieser T, Altenbuchner J (2003) *Streptomyces coelicolor* A3 (2) plasmid SCP2*: deductions from the complete sequence. Microbiology 149(Pt2):505-13. doi:10.1099/mic.0.25751-0
- Brune I, Jochmann N, Brinkrolf K, Hüser AT, Gerstmeir R, Eikmanns BJ, Kalinowski J, Pühler A, Tauch A (2007) The IclR-type transcriptional repressor LtbR regulates the expression of leucine and tryptophan biosynthesis genes in the amino acid producer *Corynebacterium glutamicum*. Journal of bacteriology 189 (7):2720-2733. doi:10.1128/JB.01876-06
- Bursy J, Kuhlmann AU, Pittelkow M, Hartmann H, Jebbar M, Pierik AJ, Bremer E (2008) Synthesis and uptake of the compatible solutes ectoine and 5-hydroxyectoine by *Streptomyces coelicolor* A3(2) in response to salt and heat stresses. Applied and environmental microbiology 74 (23):7286-7296. doi:10.1128/AEM.00768-08
- Bush MJ, Bibb MJ, Chandra G, Findlay KC, Buttner MJ (2013) Genes required for aerial growth, cell division, and chromosome segregation are targets of WhiA before sporulation in *Streptomyces venezuelae*. mBio 4 (5):13. doi:10.1128/mBio.00684-13
- Cane DE, Watt RM (2003) Expression and mechanistic analysis of a germacradienol synthase from *Streptomyces coelicolor* implicated in geosmin biosynthesis. Proceedings of the National Academy of Sciences of the United States of America 100 (4):1547-1551. doi:10.1073/pnas.0337625100
- Chatterjee A, Drews L, Mehra S, Takano E, Kaznessis YN, Hu WS (2011) Convergent transcription in the butyrolactone regulon in *Streptomyces*

- coelicolor* confers a bistable genetic switch for antibiotic biosynthesis. PLoS One 6 (7):e21974. doi:10.1371/journal.pone.0021974
- Chen L, Chen J, Jiang Y, Zhang W (2009) Transcriptomics analyses reveal global roles of the regulator AveI in *Streptomyces avermitilis*. FEMS microbiology letters 298(2):199-207 doi:10.1111/j.1574-6968.2009.01721.x
- Consortium I (2004) Finishing the euchromatic sequence of the human genome. Nature. doi:10.1038/nature03001
- Cortes T, Schubert OT, Rose G, Arnvig KB, Comas I (2013) Genome-wide mapping of transcriptional start sites defines an extensive leaderless transcriptome in *Mycobacterium tuberculosis*. Cell reports 5(4):1121-31
- Craney A, Ahmed S, Nodwell J (2013) Towards a new science of secondary metabolism. The Journal of antibiotics 66 (7):387-400. doi:10.1038/ja.2013.25
- Craster HL, Potter CA, Baumberg S (1999) End-product control of expression of branched-chain amino acid biosynthesis genes in *Streptomyces coelicolor* A3(2): paradoxical relationships between DNA sequence and regulatory phenotype. Microbiology 145 (Pt 9):2375-2384
- Crooks GE, Hon G, Chandonia JM (2004) WebLogo: a sequence logo generator. Genome research 14(6):1188-90
- De Rossi E, Leva R, Gusberti L, Manachini PL, Riccardi G (1995) Cloning, sequencing and expression of the ilvBNC gene cluster from *Streptomyces avermitilis*. Gene 166 (1):127-132
- Derek JL, Carmen M, Helen MK, David AH (1988) Mutation and cloning of clustered *Streptomyces* genes essential for sulphate metabolism. MGG

Molecular & General Genetics 211. doi:10.1007/BF00425694

- Donadio S, Shafiee A, Hutchinson CR (1990) Disruption of a rhodanese-like gene results in cysteine auxotrophy in *Saccharopolyspora erythraea*. Journal of bacteriology 172 (1):350-360
- Dosanjh NS, Rawat M, Chung J-HH, Av-Gay Y (2005) Thiol specific oxidative stress response in *Mycobacteria*. FEMS microbiology letters 249 (1):87-94. doi:10.1016/j.femsle.2005.06.004
- Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B (2003) Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proceedings of the National Academy of Sciences of the United States of America 100 (15):8817-8822. doi:10.1073/pnas.1133470100
- Dyson P (2011) *Streptomyces*: molecular biology and biotechnology. Horizon Scientific Press,
- Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S (2009) Real-time DNA sequencing from single polymerase molecules. Science 323 (5910):133-138. doi:10.1126/science.1162986

- Feng J, Liu T, Qin B, Zhang Y, Liu XS (2012) Identifying ChIP-seq enrichment using MACS. *Nature protocols* 7 (9):1728-1740. doi:10.1038/nprot.2012.101
- Feng WH, Mao XM, Liu ZH, Li YQ (2011) The ECF sigma factor SigT regulates actinorhodin production in response to nitrogen stress in *Streptomyces coelicolor*. *Applied microbiology and biotechnology* 92(5):1009-21 doi:10.1007/s00253-011-3619-2
- Fischer M, Schmidt C, Falke D, Sawers RG (2012) Terminal reduction reactions of nitrate and sulfate assimilation in *Streptomyces coelicolor* A3(2): identification of genes encoding nitrite and sulfite reductases. *Research in microbiology* 163 (5):340-348. doi:10.1016/j.resmic.2012.05.004
- Flärdh K, Buttner MJ (2009) *Streptomyces* morphogenetics: dissecting differentiation in a filamentous bacterium. *Nature Reviews Microbiology* 7 (1):36-49. doi:10.1038/nrmicro1968
- Flett F, Mersinias V, Smith CP (1997) High efficiency intergeneric conjugal transfer of plasmid DNA from *Escherichia coli* to methyl DNA-restricting streptomycetes. *FEMS microbiology letters*. 155(2):223-9 doi:10.1111/j.1574-6968.1997.tb13882.x
- Güell M, Yus E, Lluch-Senar M, Serrano L (2011) Bacterial transcriptomics: what is beyond the RNA hori-z-ome? *Nature reviews Microbiology* 9 (9):658-669. doi:10.1038/nrmicro2620
- Gilchrist DA, Fargo DC, Adelman K (2009) Using ChIP-chip and ChIP-seq to study the regulation of gene expression: Genome-wide localization studies reveal widespread regulation of transcription elongation. *Methods* 48 (4):398-408. doi:DOI 10.1016/j.ymeth.2009.02.024

- Goentoro L, Shoval O, Kirschner MW, Alon U (2009) The incoherent feedforward loop can provide fold-change detection in gene regulation. *Molecular cell* 36 (5):894-899. doi:10.1016/j.molcel.2009.11.018
- Gust B, Challis GL, Fowler K, Kieser T, Chater KF (2003) PCR-targeted *Streptomyces* gene replacement identifies a protein domain needed for biosynthesis of the sesquiterpene soil odor geosmin. *Proceedings of the National Academy of Sciences of the United States of America* 100 (4):1541-1546. doi:DOI 10.1073/pnas.0337542100
- Gust B, Chandra G, Jakimowicz D, Yuqing T, Bruton CJ, Chater KF (2004) Lambda red-mediated genetic manipulation of antibiotic-producing *Streptomyces*. *Advanced in applied microbiology* 54:107-128. doi:10.1016/S0065-2164(04)54004-2
- Hang W, Shuang Q, Chenyang L, Huajun Z, Xiufen Z, Linqun B, Zixin D (2012) Genomic and transcriptomic insights into the thermo-regulated biosynthesis of validamycin in *Streptomyces hygroscopicus* 5008. *BMC Genomics* 13. doi:10.1186/1471-2164-13-337
- Hengst DCD, Tran NT, Bibb MJ (2010) Genes essential for morphological development and antibiotic production in *Streptomyces coelicolor* are targets of BldD during vegetative growth. *Molecular microbiology* 78(2):361-79 doi:10.1111/j.1365-2958.2010.07338.x
- Hesketh A, Chen WJ, Ryding J, Chang S, Bibb M (2007) The global role of ppGpp synthesis in morphological differentiation and antibiotic production in *Streptomyces coelicolor* A3 (2). *Genome biology* 8(8):R161
- Hesketh AR, Chandra G, Shaw AD, Rowland JJ, Kell DB, Bibb MJ, Chater KF

- (2002) Primary and secondary metabolism, and post-translational protein modifications, as portrayed by proteomic analysis of *Streptomyces coelicolor*. *Molecular microbiology* 46 (4):917-932. doi:10.1046/j.1365-2958.2002.03219.x
- Higo A, Hara H, Horinouchi S, Ohnishi Y (2012) Genome-wide distribution of AdpA, a global regulator for secondary metabolism and morphological differentiation in *Streptomyces*, revealed the extent and complexity of the AdpA regulatory network. *DNA research : an international journal for rapid publication of reports on genes and genomes* 19 (3):259-273. doi:10.1093/dnares/dss010
- Huang J, Lih CJ, Pan KH, Cohen SN (2001) Global analysis of growth phase responsive gene expression and regulation of antibiotic biosynthetic pathways in *Streptomyces coelicolor* using DNA microarrays. *Genes & development* 15 (23):3183-3192. doi:10.1101/gad.943401
- Hunkapiller T, Kaiser RJ, Koop BF, Hood L (1991) Large-scale and automated DNA sequence determination. *Science* 254 (5028):59-67
- Jayapal KP, Lian W, Glod F, Sherman DH, Hu W-SS (2007) Comparative genomic hybridizations reveal absence of large *Streptomyces coelicolor* genomic islands in *Streptomyces lividans*. *BMC genomics* 8:229. doi:10.1186/1471-2164-8-229
- Joshi V, Joung J-GG, Fei Z, Jander G (2010) Interdependence of threonine, methionine and isoleucine metabolism in plants: accumulation and transcriptional regulation under abiotic stress. *Amino acids* 39 (4):933-947. doi:10.1007/s00726-010-0505-7

- Kang SH, Huang J, Lee HN, Hur YA (2007) Interspecies DNA microarray analysis identifies WblA as a pleiotropic down-regulator of antibiotic biosynthesis in *Streptomyces*. *Journal of bacteriology* 189(11):4315-9 doi:10.1128/JB.01789-06
- Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S, Chen F, Lapidus A, Ferriera S, Johnson J, Steglich C, Church GM, Richardson P, Chisholm SW (2007) Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS genetics* 3(12):e231. doi:10.1371/journal.pgen.0030231
- Kim DW, Chater K, Lee KJ, Hesketh A (2005) Changes in the extracellular proteome caused by the absence of the bldA gene product, a developmentally significant tRNA, reveal a new target for the pleiotropic regulator AdpA in *Streptomyces coelicolor*. *Journal of bacteriology* 187(9):2957-66 doi:10.1128/JB.187.9.2957-2966.2005
- Kim M-S, Dufour YS, Yoo JS, Cho Y-B, Park J-H, Nam G-B, Kim HM, Lee K-L, Donohue TJ, Roe J-H (2012a) Conservation of thiol-oxidative stress responses regulated by SigR orthologues in actinomycetes. *Molecular microbiology* 85 (2):326-344. doi:10.1111/j.1365-2958.2012.08115.x
- Kim SH, Lee B-R, Kim J-N, Kim B-G (2012b) NdgR, a common transcriptional activator for methionine and leucine biosynthesis in *Streptomyces coelicolor*. *Journal of bacteriology* 194 (24):6837-6846. doi:10.1128/JB.00695-12
- Komatsu M, Uchiyama T, Omura S, Cane DE, Ikeda H (2010) Genome-minimized *Streptomyces* host for the heterologous expression of secondary metabolism. *Proceedings of the National Academy of Sciences of the United States of*

- America 107 (6):2646-2651. doi:10.1073/pnas.0914833107
- Kopecky J, Janata J, Pospisil S, Felsberg J, Spizek J (1999) Mutations in two distinct regions of acetolactate synthase regulatory subunit from *Streptomyces cinnamonensis* result in the lack of sensitivity to end-product inhibition. Biochemical and biophysical research communications 266 (1):162-166. doi:10.1006/bbrc.1999.1792
- Lee E-J, Karoonuthaisiri N, Kim H-S, Park J-H, Cha C-J, Kao CM, Roe J-H (2005) A master regulator sigmaB governs osmotic and oxidative response as well as differentiation via a network of sigma factors in *Streptomyces coelicolor*. Molecular microbiology 57 (5):1252-1264. doi:10.1111/j.1365-2958.2005.04761.x
- Lian W, Jayapal KP, Charaniya S, Mehra S, Glod F, Kyung Y-SS, Sherman DH, Hu W-S (2008) Genome-wide transcriptome analysis reveals that a pleiotropic antibiotic regulator, AfsS, modulates nutritional stress response in *Streptomyces coelicolor* A3(2). BMC genomics 9:56. doi:10.1186/1471-2164-9-56
- Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD (2008) Highly Integrated Single-Base Resolution Maps of the Epigenome in *Arabidopsis*. Cell 133(3):523-36
- Lynda JD, David JH, Susan LC, Werner E, Kenneth GS, Harry WD (2008) Mass spectrometric study of the *Escherichia coli* repressor proteins, IcIR and GcIR, and their complexes with DNA. Protein Science 10(7):1370-80. doi:10.1110/ps.780101
- Mangan S, Alon U (2003) Structure and function of the feed-forward loop network

- motif. Proceedings of the National Academy of Sciences 100.
doi:10.1073/pnas.2133841100
- Mardis ER (2008) Next-generation DNA sequencing methods. Annual review of
genomics and human genetics 9:387-402.
doi:10.1146/annurev.genom.9.081307.164359
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J,
Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV,
Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML,
Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz
SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna
MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth
GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz
A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF,
Rothberg JM (2005) Genome sequencing in microfabricated high-density
picolitre reactors. Nature 437 (7057):376-380. doi:10.1038/nature03959
- Medini D, Donati C, Tettelin H, Massignani V (2005) The microbial pan-genome.
Current opinion in genetics & development 15(6):589-94
doi:10.1016/j.gde.2005.09.006
- Metzker ML (2009) Sequencing technologies - the next generation. Nature Reviews
Genetics 11(1):31-46. doi:10.1038/nrg2626
- Mitra RD, Church GM (1999) In situ localized amplification and contact replication
of many individual DNA molecules. Nucleic Acids Res 27 (24):e34
- Mitra RD, Shendure J, Olejnik J, Edyta Krzymanska O, Church GM (2003)
Fluorescent in situ sequencing on polymerase colonies. Analytical

biochemistry 320 (1):55-65

- Moody MJ, Young RA, Jones SE, Elliot MA (2013) Comparative analysis of non-coding RNAs in the antibiotic-producing *Streptomyces bacteria*. BMC Genomics 14:558. doi:10.1186/1471-2164-14-558
- Nègre D, Cortay JC, Old IG, Galinier A, Richaud C, Saint Girons I, Cozzzone AJ (1991) Overproduction and characterization of the iclR gene product of Escherichia coli K-12 and comparison with that of *Salmonella typhimurium* LT2. Gene 97 (1):29-37
- Nai-Hua H, Ralph K (2007) Comparative genomics of *Streptomyces avermitilis*, *Streptomyces cattleya*, *Streptomyces maritimus* and *Kitasatospora aureofaciens* using a *Streptomyces coelicolor* microarray system. Antonie van Leeuwenhoek. 93(1-2):1-25 doi:10.1007/s10482-007-9175-1
- Nestor Z, Mariia R, Bohdan O, Victor F, Andriy L (2014) Insights into naturally minimised *Streptomyces albus* J1074 genome. BMC Genomics. doi:10.1186/1471-2164-15-97
- Nett M, Ikeda H, Moore BS (2009) Genomic basis for natural product biosynthetic diversity in the actinomycetes. Natural product reports 26 (11):1362-1384. doi:10.1039/b817069j
- Novotna J, Vohradsky J, Berndt P (2003) Proteomic studies of diauxic lag in the differentiating prokaryote *Streptomyces coelicolor* reveal a regulatory network of stress-induced proteins and central metabolic enzymes Molecular microbiology 48(5):1289-303. doi:10.1046/j.1365-2958.2003.03529.x
- Obata T, Fernie AR (2012) The use of metabolomics to dissect plant responses to

- abiotic stresses. Cellular and molecular life sciences : CMLS 69 (19):3225-3243. doi:10.1007/s00018-012-1091-5
- Ohnishi Y, Ishikawa J, Hara H, Suzuki H, Ikenoya M, Ikeda H, Yamashita A, Hattori M, Horinouchi S (2008) Genome sequence of the streptomycin-producing microorganism *Streptomyces griseus* IFO 13350. Journal of bacteriology 190 (11):4050-4060. doi:10.1128/JB.00204-08
- Orlando V (2000) Mapping chromosomal proteins *in vivo* by formaldehyde-crosslinked-chromatin immunoprecipitation. Trends in biochemical sciences 25(3):99-104
- Pérez-Redondo R, Rodríguez-García A, Botas A, Santamarta I, Martín JF, Liras P (2012) ArgR of *Streptomyces coelicolor* is a versatile regulator. PloS one 7 (3). doi:10.1371/journal.pone.0032697
- Paget MS, Molle V, Cohen G, Aharonowitz Y, Buttner MJ (2001) Defining the disulphide stress response in *Streptomyces coelicolor* A3(2): identification of the sigmaR regulon. Molecular microbiology 42 (4):1007-1020
- Pan B, Unnikrishnan I, LaPorte DC (1996) The binding site of the IclR repressor protein overlaps the promoter of aceBAK. Journal of bacteriology 178 (13):3982-3984
- Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. Nature reviews Genetics 10 (10):669-680. doi:10.1038/nrg2641
- Poralla K, Muth G, Hartner T (2000) Hopanoids are formed during transition from substrate to aerial hyphae in *Streptomyces coelicolor* A3(2). FEMS Microbiology Letters 189 (1):93-95
- Pullan ST, Chandra G, Bibb MJ, Merrick M (2011) Genome-wide analysis of the

role of GlnR in *Streptomyces venezuelae* provides new insights into global nitrogen regulation in actinomycetes. BMC genomics

- Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13:341. doi:10.1186/1471-2164-13-341
- Rach EA, Yuan HY, Majoros WH, Tomancak P, Ohler U (2009) Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the *Drosophila* genome. Genome Biol 10 (7):R73. doi:10.1186/gb-2009-10-7-r73
- Rasko DA, Rosovitz MJ, Myers GSA, Mongodin EF, Fricke FW, Gajer P, Crabtree J, Sebaihia M, Thomson NR, Chaudhuri R (2008) The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. Journal of bacteriology 190 (20):6881-6893. doi:10.1128/JB.00619-08
- Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG (2000) Genome-wide location and function of DNA binding proteins. Science. doi:10.1126/science.290.5500.2306
- Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, Jones S (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nature methods 4(8):651-7 doi:10.1038/nmeth1068
- Romero DA, Hasan AH, Lin Y-f, Kime L, Ruiz-Larrabeiti O, Urem M, Bucca G,

- Mamanova L, Laing EE, van Wezel GP, Smith CP, Kaberdin VR, McDowall KJ (2014) A comparison of key aspects of gene regulation in *Streptomyces coelicolor* and *Escherichia coli* using nucleotide-resolution transcription maps produced in parallel by global and differential RNA sequencing. *Molecular Microbiology*:n/a-n/a. doi:10.1111/mmi.12810
- Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Analytical biochemistry* 242 (1):84-89. doi:10.1006/abio.1996.0432
- Rotem Sorek PC (2009) Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nature Reviews Genetics* 11 (1):9-16. doi:10.1038/nrg2695
- Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, Hoon J, Simons JF, Marran D, Myers JW, Davidson JF, Branting A, Nobile JR, Puc BP, Light D, Clark TA, Huber M, Branciforte JT, Stoner IB, Cawley SE, Lyons M, Fu Y, Homer N, Sedova M, Miao X, Reed B, Sabina J, Feierstein E, Schorn M, Alanjary M, Dimalanta E, Dressman D, Kasinskas R, Sokolsky T, Fidanza JA, Namsaraev E, McKernan KJ, Williams A, Roth GT, Bustillo J (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 475 (7356):348-352. doi:10.1038/nature10242
- Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M (1977) Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 265 (5596):687-695
- Santamarta I, López-García MT, Pérez-Redondo R, Koekman B, Martín JF, Liras P

- (2007) Connecting primary and secondary metabolism: AreB, an IclR-like protein, binds the ARE(ccaR) sequence of *S. clavuligerus* and modulates leucine biosynthesis and cephamycin C and clavulanic acid production. *Molecular microbiology* 66 (2):511-524. doi:10.1111/j.1365-2958.2007.05937.x
- Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*. doi:10.1126/science.270.5235.467
- Selinger DW, Cheung KJ, Mei R, Johansson EM (2000) RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome array. *Nature biotechnology* 18(12):1262-8 doi:10.1038/82367
- Seliverstov AV, Putzer H, Gelfand MS, Lyubetsky VA (2005) Comparative analysis of RNA regulatory elements of amino acid metabolism genes in Actinobacteria. *BMC microbiology* 5:54. doi:10.1186/1471-2180-5-54
- Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, Chabas S, Reiche K, Hackermüller J, Reinhardt R, Stadler PF, Vogel J (2010) The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* 464 (7286):250-255. doi:10.1038/nature08756
- Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nature biotechnology* 26 (10):1135-1145. doi:10.1038/nbt1486
- Spencer VA, Sun JM, Li L, Davie JR (2003) Chromatin immunoprecipitation: a tool for studying histone acetylation and transcription factor binding. *Methods* 31 (1):67-75. doi:S1046202303000896
- Stroupe ME, Leech HK, Daniels DS, Warren MJ, Getzoff ED (2003) CysG structure

- reveals tetrapyrrole-binding features and novel regulation of siroheme biosynthesis. *Nature structural biology* 10 (12):1064-1073. doi:10.1038/nsb1007
- Sugawara M, Epstein B, Badgley BD, Unno... T (2013) Comparative genomics of the core and accessory genomes of 48 *Sinorhizobium* strains comprising five genospecies. *Genome biology* 14(2):R17 doi:10.1186/gb-2013-14-2-r17
- Swerdlow H, Wu SL, Harke H, Dovichi NJ (1990) Capillary gel electrophoresis for DNA sequencing. Laser-induced fluorescence detection with the sheath flow cuvette. *Journal of chromatography* 516 (1):61-67
- Takano E, Kinoshita H, Mersinias V, Bucca G, Hotchkiss G, Nihira T, Smith CP, Bibb M, Wohlleben W, Chater K (2005) A bacterial hormone (the SCB1) directly controls the expression of a pathway-specific regulatory gene in the cryptic type I polyketide biosynthetic gene cluster of *Streptomyces coelicolor*. *Mol Microbiol* 56 (2):465-479. doi:MMI4543 10.1111/j.1365-2958.2005.04543.x
- Tatusov RL, Natale DA, Garkavtsev IV (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic acids research* 29(1):22-8 doi:10.1093/nar/29.1.22
- Tettelin H, Massignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJ, Smith S, Utterback TR,

- White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM. (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”. *Proceedings of the National Academy of Sciences of the United States of America* 102(39):13950-5 doi:10.1073/pnas.0506758102
- Viollier PH, Kelemen GH, Dale GE (2003) Specialized osmotic stress response systems involve multiple SigB-like sigma factors in *Streptomyces coelicolor*. *Molecular microbiology* 47(3):699-714. doi:10.1046/j.1365-2958.2003.03302.x
- Vockenhuber MP, Sharma CM, Statt MG, Schmidt D, Xu Z, Dietrich S, Liesegang H, Mathews DH, Suess B (2011) Deep sequencing-based identification of small non-coding RNAs in *Streptomyces coelicolor*. *RNA biology* 8 (3):468-477
- Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*. 10(1):57-63 doi:10.1038/nrg2484
- Yang Y-H, Song E, Kim E-J, Lee K, Kim W-S, Park S-S, Hahn J-S, Kim B-G (2009a) NdgR, an IclR-like regulator involved in amino-acid-dependent growth, quorum sensing, and antibiotic production in *Streptomyces coelicolor*. *Applied microbiology and biotechnology* 82 (3):501-511. doi:10.1007/s00253-008-1802-x
- Zhao Y, Wu J, Yang J, Sun S, Xiao J, Yu J (2012) PGAP: pan-genomes analysis pipeline. *Bioinformatics*. 28(3):416-8 doi:10.1093/bioinformatics/btr655

- Zheng X, Hu GQ, She ZS, Zhu H (2011) Leaderless genes in bacteria: clue to the evolution of translation initiation mechanisms in prokaryotes. BMC Genomics 12:361. doi:10.1186/1471-2164-12-361
- Zhou Z, Gu J, Li Y-Q, Wang Y (2012) Genome plasticity and systems evolution in *Streptomyces*. BMC bioinformatics 13 (Suppl 10). doi:10.1186/1471-2105-13-S10-S8

APPENDIX

Appendix I The list of leaderless genes in *S. coelicolor*.

SCO No.	Strand	position	UTR length	SCO No.	Strand	position	UTR length
SCO0055	-	43682	0	SCO0884	-	931212	0
SCO0142	-	136100	0	SCO0908	+	952991	0
SCO0155	+	145499	0	SCO0916	-	961283	0
SCO0185	+	173768	0	SCO0917	+	961534	0
SCO0214	+	204387	0	SCO0931	+	978713	0
SCO0222	+	214913	0	SCO0937	+	983389	0
SCO0224	-	216844	0	SCO0945	-	991112	0
SCO0241	+	229995	0	SCO0946	-	993243	0
SCO0244	+	232508	0	SCO0954	-	1004478	0
SCO0253	+	242194	0	SCO1013	-	1069633	0
SCO0262	+	250805	0	SCO1017	-	1073219	0
SCO0277	-	269688	0	SCO1019	+	1074142	0
SCO0278	+	269783	0	SCO1022	+	1077501	0
SCO0298	-	293519	0	SCO1039	-	1096518	0
SCO0302	-	298108	0	SCO1092	+	1151356	0
SCO0316	-	316921	0	SCO1094	+	1152798	0
SCO0337	+	340752	0	SCO1102	+	1160123	0
SCO0405	-	426833	0	SCO1114	+	1171863	0
SCO0407	+	427610	0	SCO1122	-	1180792	0
SCO0426	+	444441	0	SCO1123	+	1180851	0
SCO0441	+	461350	0	SCO1145	+	1204428	0
SCO0447	+	465990	0	SCO1156	+	1216584	0
SCO0463	+	484318	0	SCO1162	+	1222188	0
SCO0475	-	497359	0	SCO1178	-	1242783	0
SCO0485	+	504854	0	SCO1189	+	1262121	0
SCO0513	+	547809	0	SCO1200	-	1272566	0
SCO0521	+	553892	0	SCO1220	-	1292117	0
SCO0558	-	599978	0	SCO1221	+	1292195	0
SCO0589	-	632103	0	SCO1244	+	1316289	0
SCO0605	+	645297	0	SCO1250	+	1321315	0
SCO0637	+	678937	0	SCO1254	+	1325285	0
SCO0646	+	688679	0	SCO1260	+	1330916	0
SCO0722	-	767331	0	SCO1263	-	1334711	0
SCO0731	-	774613	0	SCO1295	+	1370221	0
SCO0738	+	780863	0	SCO1296	+	1370791	0
SCO0747	-	789837	0	SCO1297	-	1373469	0
SCO0802	-	850042	0	SCO1302	-	1379752	0
SCO0822	-	872086	0	SCO1308	+	1385454	0
SCO0848	-	895819	0	SCO1311	-	1388832	0
SCO0858	-	905377	0	SCO1318	+	1393213	0
SCO0877	+	922247	0	SCO1319	+	1394273	0

SCO1320	-	1396164	0	SCO1789	+	1913090	0
SCO1322	-	1398413	0	SCO1797	-	1926662	0
SCO1334	+	1411647	0	SCO1799	+	1927668	0
SCO1340	+	1417147	0	SCO1802	-	1930823	0
SCO1358	-	1436223	0	SCO1828	+	1959189	0
SCO1359	+	1436298	0	SCO1830	+	1960424	0
SCO1360	-	1439097	0	SCO1831	+	1962217	0
SCO1364	-	1442748	0	SCO1861	-	1996295	0
SCO1371	+	1447950	0	SCO1870	-	2004454	0
SCO1382	-	1461209	0	SCO1873	+	2007027	0
SCO1410	-	1505123	0	SCO1878	-	2011522	0
SCO1419	+	1514933	0	SCO1916	-	2051513	0
SCO1422	-	1518445	0	SCO1926	-	2059132	0
SCO1447	-	1544653	0	SCO1928	+	2060169	0
SCO1464	-	1562157	0	SCO1953	-	2089856	0
SCO1469	-	1570513	0	SCO1959	+	2097542	0
SCO1475	-	1577602	0	SCO1960	+	2098289	0
SCO1479	-	1581474	0	SCO1966	-	2107199	0
SCO1530	-	1636876	0	SCO1969	+	2109316	0
SCO1533	-	1640971	0	SCO1973	-	2113110	0
SCO1535	-	1642515	0	SCO1975	-	2115132	0
SCO1556	-	1668945	0	SCO1989	-	2127773	0
SCO1561	+	1672877	0	SCO1996	-	2131993	0
SCO1571	+	1681979	0	SCO2003	-	2144382	0
SCO1582	-	1692572	0	SCO2014	-	2157773	0
SCO1602	-	1713279	0	SCO2043	-	2193732	0
SCO1609	+	1720261	0	SCO2045	+	2194212	0
SCO1616	-	1729404	0	SCO2057	+	2205747	0
SCO1640	-	1755251	0	SCO2070	+	2220881	0
SCO1651	-	1766159	0	SCO2073	-	2225155	0
SCO1663	-	1784336	0	SCO2100	+	2257287	0
SCO1670	+	1789862	0	SCO2103	-	2261916	0
SCO1672	-	1794029	0	SCO2105	-	2263246	0
SCO1678	-	1798424	0	SCO2112	-	2270952	0
SCO1687	+	1807505	0	SCO2115	-	2273425	0
SCO1702	+	1821800	0	SCO2125	-	2285838	0
SCO1714	-	1835573	0	SCO2157	+	2319777	0
SCO1719	+	1839862	0	SCO2158	-	2322298	0
SCO1727	+	1847564	0	SCO2175	+	2338274	0
SCO1729	+	1848849	0	SCO2196	+	2363402	0
SCO1733	-	1851669	0	SCO2197	-	2364684	0
SCO1753	+	1874625	0	SCO2201	+	2367728	0
SCO1754	-	1875975	0	SCO2211	+	2375406	0
SCO1764	+	1884039	0	SCO2212	+	2376423	0

SCO2223	+	2387313	0	SCO2670	-	2907651	0
SCO2233	+	2401780	0	SCO2755	-	3002577	0
SCO2243	+	2412355	0	SCO2765	+	3015897	0
SCO2266	+	2434908	0	SCO2775	-	3027416	0
SCO2267	+	2435930	0	SCO2787	+	3042474	0
SCO2268	-	2437237	0	SCO2788	+	3043459	0
SCO2287	+	2457954	0	SCO2793	+	3048673	0
SCO2301	-	2463312	0	SCO2847	+	3106290	0
SCO2295	+	2465169	0	SCO2860	-	3116931	0
SCO2301	-	2471173	0	SCO2893	-	3147951	0
SCO2308	+	2477055	0	SCO2899	-	3154226	0
SCO2315	+	2486429	0	SCO2901	+	3154677	0
SCO2328	+	2498914	0	SCO2902	-	3155892	0
SCO2340	+	2510090	0	SCO2904	-	3157238	0
SCO2343	+	2511913	0	SCO2927	-	3179765	0
SCO2354	+	2522814	0	SCO2934	-	3186324	0
SCO2357	+	2525802	0	SCO2938	+	3191138	0
SCO2364	+	2533506	0	SCO2972	-	3226936	0
SCO2373	-	2544617	0	SCO3000	-	3233271	0
SCO2376	+	2546758	0	SCO3006	+	3274746	0
SCO2386	+	2558045	0	SCO3014	-	3280002	0
SCO2387	+	2559340	0	SCO3036	+	3290726	0
SCO2393	+	2564900	0	SCO3063	-	3322266	0
SCO2395	-	2568256	0	SCO3119	-	3356667	0
SCO2398	+	2570030	0	SCO3133	+	3421340	0
SCO2462	+	2647782	0	SCO3173	+	3434529	0
SCO2468	-	2658056	0	SCO3177	-	3477824	0
SCO2473	+	2662608	0	SCO3182	-	3483236	0
SCO2483	+	2673003	0	SCO3188	+	3487461	0
SCO2486	+	2676046	0	SCO3199	-	3494975	0
SCO2489	+	2680340	0	SCO3205	+	3507907	0
SCO2505	+	2704004	0	SCO3224	-	3513024	0
SCO2516	+	2711253	0	SCO3269	+	3536808	0
SCO2529	+	2727564	0	SCO3273	+	3615480	0
SCO2531	+	2728980	0	SCO3297	+	3618213	0
SCO2544	+	2742662	0	SCO3311	-	3646131	0
SCO2557	+	2756986	0	SCO3329	-	3663042	0
SCO2569	-	2774499	0	SCO3333	-	3681213	0
SCO2605	+	2826447	0	SCO3348	-	3685341	0
SCO2615	-	2841581	0	SCO3349	+	3707327	0
SCO2641	+	2871981	0	SCO3351	+	3707399	0
SCO2644	-	2877352	0	SCO3361	-	3710688	0
SCO2647	+	2879447	0	SCO3363	+	3720577	0
SCO2648	-	2880369	0			3721342	0

SCO3369	+	3727304	0	SCO3953	-	4353010	0
SCO3376	-	3741220	0	SCO3961	-	4361825	0
SCO3387	+	3750688	0	SCO3962	-	4363267	0
SCO3390	-	3754484	0	SCO3976	-	4379098	0
SCO3391	+	3754603	0	SCO3979	+	4382030	0
SCO3392	-	3756807	0	SCO3981	-	4384371	0
SCO3398	+	3764143	0	SCO4015	+	4410971	0
SCO3407	-	3773342	0	SCO4016	+	4412004	0
SCO3419	-	3785358	0	SCO4019	+	4414377	0
SCO3433	+	3793755	0	SCO4024	+	4421136	0
SCO3434	+	3794733	0	SCO4047	-	4439791	0
SCO3435	+	3796515	0	SCO4038	-	4456952	0
SCO3559	-	3935802	0	SCO4067	+	4457168	0
SCO3566	+	3945434	0	SCO4075	-	4469743	0
SCO3577	+	3954997	0	SCO4094	+	4492828	0
SCO3580	+	3958243	0	SCO4097	+	4496497	0
SCO3581	-	3961105	0	SCO4109	+	4508515	0
SCO3582	+	3961226	0	SCO4122	+	4529468	0
SCO3590	-	3968617	0	SCO4145	-	4562289	0
SCO3610	-	3989856	0	SCO4158	-	4576134	0
SCO3622	+	4001665	0	SCO4180	+	4590561	0
SCO3633	+	4010391	0	SCO4181	+	4591019	0
SCO3650	-	4029494	0	SCO4186	+	4595605	0
SCO3658	+	4036689	0	SCO4192	+	4600336	0
SCO3678	-	4061654	0	SCO4197	+	4606581	0
SCO3693	+	4074209	0	SCO4198	+	4607362	0
SCO3710	+	4086039	0	SCO4205	+	4614619	0
SCO3714	-	4090508	0	SCO4206	-	4615969	0
SCO3773	+	4147062	0	SCO4207	+	4616019	0
SCO3792	+	4169611	0	SCO4209	+	4618020	0
SCO3796	+	4174717	0	SCO4216	-	4624796	0
SCO3801	+	4181609	0	SCO4229	+	4633199	0
SCO3805	+	4184790	0	SCO4237	+	4641923	0
SCO3810	-	4189715	0	SCO4238	-	4644464	0
SCO3818	+	4197520	0	SCO4256	+	4667076	0
SCO3822	-	4203092	0	SCO4257	+	4668591	0
SCO3840	-	4223718	0	SCO4258	+	4670132	0
SCO3864	+	4248804	0	SCO4265	-	4681806	0
SCO3871	+	4258993	0	SCO4269	-	4685459	0
SCO3898	-	4293666	0	SCO4298	+	4714155	0
SCO3912	+	4308427	0	SCO4300	-	4716272	0
SCO3913	-	4310339	0	SCO4303	-	4718222	0
SCO3917	-	4313757	0	SCO4309	+	4722391	0
SCO3918	-	4314318	0	SCO4311	-	4724157	0

SCO4321	-	4734673	0	SCO4824	+	5253981	0
SCO4324	-	4737265	0	SCO4835	-	5267634	0
SCO4326	-	4738692	0	SCO4848	+	5279853	0
SCO4332	-	4748436	0	SCO4850	-	5282203	0
SCO4336	+	4750828	0	SCO4871	+	5300705	0
SCO4352	+	4767684	0	SCO4872	-	5302356	0
SCO4353	-	4768972	0	SCO4878	-	5310356	0
SCO4358	-	4773021	0	SCO4879	+	5310555	0
SCO4364	+	4777590	0	SCO4896	-	5330840	0
SCO4367	+	4779458	0	SCO4897	+	5330904	0
SCO4368	-	4783492	0	SCO4898	-	5332572	0
SCO4389	+	4805897	0	SCO4900	-	5334410	0
SCO4398	-	4816163	0	SCO4901	+	5334546	0
SCO4403	+	4821748	0	SCO4904	-	5337183	0
SCO4415	+	4833116	0	SCO4907	-	5340202	0
SCO4416	-	4835376	0	SCO4915	+	5348311	0
SCO4422	+	4838327	0	SCO4917	-	5351948	0
SCO4431	-	4851852	0	SCO4927	+	5360721	0
SCO4434	+	4853069	0	SCO4944	-	5378730	0
SCO4436	+	4855250	0	SCO4957	-	5393750	0
SCO4450	+	4872011	0	SCO4962	+	5396674	0
SCO4454	-	4876625	0	SCO4968	+	5402117	0
SCO4457	-	4880076	0	SCO4984	-	5422911	0
SCO4490	+	4909614	0	SCO4992	+	5431032	0
SCO4496	+	4915292	0	SCO5024	+	5458019	0
SCO4499	+	4917934	0	SCO5036	-	5473182	0
SCO4506	+	4926246	0	SCO5042	-	5482009	0
SCO4556	+	4974767	0	SCO5055	-	5494679	0
SCO4590	-	5013400	0	SCO5059	+	5499055	0
SCO4631	-	5054666	0	SCO5071	-	5514249	0
SCO4633	-	5059681	0	SCO5072	+	5514323	0
SCO4640	+	5064405	0	SCO5139	+	5586476	0
SCO4644	-	5069623	0	SCO5152	-	5600209	0
SCO4645	-	5070975	0	SCO5185	+	5641824	0
SCO4677	-	5108811	0	SCO5201	+	5658833	0
SCO4687	-	5116320	0	SCO5202	+	5659650	0
SCO4731	+	5144190	0	SCO5206	-	5667360	0
SCO4732	-	5145963	0	SCO5208	-	5668887	0
SCO4740	+	5153334	0	SCO5214	+	5673488	0
SCO4744	-	5158372	0	SCO5224	-	5684465	0
SCO4750	+	5162784	0	SCO5228	-	5689000	0
SCO4757	+	5168216	0	SCO5239	+	5700131	0
SCO4766	+	5178021	0	SCO5241	-	5703401	0
SCO4791	+	5212113	0	SCO5245	-	5706013	0

SCO5247	-	5708624	0	SCO5787	+	6324453	0
SCO5253	+	5713933	0	SCO5791	+	6327656	0
SCO5262	+	5721695	0	SCO5793	+	6329272	0
SCO5300	-	5773118	0	SCO5815	+	6359074	0
SCO5311	-	5783972	0	SCO5819	+	6366349	0
SCO5329	+	5798773	0	SCO5830	-	6381357	0
SCO5336	-	5809892	0	SCO5843	+	6397754	0
SCO5351	+	5818975	0	SCO5854	+	6410064	0
SCO5381	-	5850686	0	SCO5859	+	6416236	0
SCO5384	+	5852346	0	SCO5900	+	6465627	0
SCO5398	-	5870463	0	SCO5908	+	6474517	0
SCO5413	+	5882733	0	SCO5930	+	6498298	0
SCO5415	+	5883756	0	SCO5971	+	6542306	0
SCO5423	-	5894405	0	SCO5986	+	6562040	0
SCO5435	-	5909691	0	SCO5998	-	6574715	0
SCO5437	+	5911410	0	SCO6018	+	6600024	0
SCO5459	+	5946399	0	SCO6030	-	6619849	0
SCO5483	-	5971214	0	SCO6041	+	6632494	0
SCO5487	-	5973862	0	SCO6042	+	6633983	0
SCO5504	+	5991461	0	SCO6049	-	6641344	0
SCO5510	+	5998409	0	SCO6073	+	6666219	0
SCO5517	+	6010146	0	SCO6079	+	6674122	0
SCO5522	+	6015259	0	SCO6092	+	6692660	0
SCO5535	+	6031665	0	SCO6103	-	6703541	0
SCO5549	+	6048321	0	SCO6106	-	6708353	0
SCO5552	-	6051815	0	SCO6119	+	6724307	0
SCO5555	+	6054328	0	SCO6120	+	6725136	0
SCO5557	-	6055700	0	SCO6121	-	6726270	0
SCO5560	+	6058385	0	SCO6124	+	6727703	0
SCO5576	+	6073650	0	SCO6139	+	6740466	0
SCO5577	+	6074834	0	SCO6147	-	6748448	0
SCO5590	+	6096339	0	SCO6155	+	6756908	0
SCO5605	+	6107817	0	SCO6225	-	6847252	0
SCO5649	+	6148202	0	SCO6228	+	6850590	0
SCO5651	-	6151109	0	SCO6253	+	6875083	0
SCO5656	-	6155949	0	SCO6254	+	6876381	0
SCO5694	+	6203922	0	SCO6255	-	6878175	0
SCO5698	+	6209024	0	SCO6283	+	6942661	0
SCO5744	+	6266753	0	SCO6339	+	6998012	0
SCO5752	+	6290145	0	SCO6340	+	6999405	0
SCO5762	+	6301456	0	SCO6341	+	7000095	0
SCO5763	+	6302319	0	SCO6384	-	7051151	0
SCO5766	+	6304454	0	SCO6390	+	7056189	0
SCO5784	+	6321291	0	SCO6409	-	7077213	0

SCO6410	+	7077312	0	SCO7244	+	8053492	0
SCO6425	+	7094757	0	SCO7249	+	8057947	0
SCO6439	-	7122904	0	SCO7269	-	8079047	0
SCO6445	+	7129505	0	SCO7270	-	8079972	0
SCO6459	-	7147390	0	SCO7286	+	8092023	0
SCO6464	-	7153332	0	SCO7293	-	8099116	0
SCO6476	+	7166743	0	SCO7366	+	8178701	0
SCO6486	+	7178603	0	SCO7400	-	8214151	0
SCO6493	-	7187208	0	SCO7427	+	8241835	0
SCO6512	+	7202065	0	SCO7433	+	8246125	0
SCO6514	+	7204898	0	SCO7434	+	8247219	0
SCO6519	+	7209379	0	SCO7438	+	8252589	0
SCO6592	-	7306619	0	SCO7440	-	8255261	0
SCO6620	+	7340945	0	SCO7453	-	8269446	0
SCO6631	+	7358814	0	SCO7455	+	8270459	0
SCO6639	+	7372859	0	SCO7458	+	8272128	0
SCO6707	-	7459941	0	SCO7478	-	8292117	0
SCO6709	+	7460584	0	SCO7510	-	8325307	0
SCO6715	+	7469875	0	SCO7514	+	8328549	0
SCO6725	-	7480202	0	SCO7519	-	8334732	0
SCO6728	+	7481296	0	SCO7520	-	8336018	0
SCO6735	+	7490061	0	SCO7539	-	8360931	0
SCO6740	-	7496600	0	SCO7562	-	8385997	0
SCO6812	+	7574420	0	SCO7605	-	8433208	0
SCO6830	+	7601654	0	SCO7613	-	8442160	0
SCO6836	+	7606744	0	SCO7634	+	8461956	0
SCO6952	-	7716844	0	SCO7639	+	8468385	0
SCO6960	-	7727039	0	SCO7694	+	8531016	0
SCO6971	-	7739795	0	SCO7696	+	8532086	0
SCO6990	-	7758473	0	SCO7702	+	8539362	0
SCO7016	-	7798381	0	SCO7704	+	8541365	0
SCO7024	+	7810502	0	SCO7706	-	8544237	0
SCO7026	+	7812182	0	SCO7709	+	8545444	0
SCO7036	+	7824771	0	SCO7715	+	8550431	0
SCO7042	-	7835298	0	SCO7721	+	8555357	0
SCO7047	+	7837144	0	SCO7727	-	8563058	0
SCO7067	+	7857217	0	SCO7731	+	8565486	0
SCO7142	+	7936071	0	SCO7732	+	8566732	0
SCO7146	+	7941058	0	SCO7741	-	8577372	0
SCO7191	-	7995370	0	SCO7742	+	8577403	0
SCO7194	+	7997836	0	SCO7743	-	8578485	0
SCO7230	+	8035811	0	SCO7767	-	8597410	0
SCO7238	+	8046083	0	SCO7772	-	8600881	0
SCO7239	+	8046501	0	SCO7773	-	8601808	0

SCO7794	+	8618313	0	SCO3140	+	3441592	3
SCO7813	+	8631557	0	SCO3294	-	3644166	3
SCO0058	+	44252	1	SCO3355	+	3714375	3
SCO1006	-	1061553	1	SCO3812	+	4190996	3
SCO1673	+	1794076	1	SCO3943	-	4339149	3
SCO1731	-	1850888	1	SCO4111	-	4511703	3
SCO1784	+	1908437	1	SCO5645	+	6143016	3
SCO1990	-	2128542	1	SCO5809	+	6352816	3
SCO2375	+	2545470	1	SCO5812	+	6356446	3
SCO4449	-	4871942	1	SCO5929	-	6498234	3
SCO4769	+	5180619	1	SCO6350	-	7012791	3
SCO4797	+	5218623	1	SCO6742	+	7497715	3
SCO5332	+	5805822	1	SCO7801	+	8621517	3
SCO5451	-	5937459	1	SCO0574	+	616249	4
SCO5831	-	6382678	1	SCO1736	-	1854528	4
SCO6260	+	6882241	1	SCO1944	-	2077945	4
SCO7604	+	8430894	1	SCO3069	-	3362034	5
SCO0420	-	438289	2	SCO4048	-	4440547	5
SCO0635	+	676128	2	SCOt53	-	5666898	5
SCO1237	-	1311106	2	SCO6623	+	7345056	5
SCO1540	-	1650187	2	SCO7649	-	8477251	5
SCO1697	-	1818880	2	SCO1268	-	1339372	6
SCO1792	-	1922771	2	SCOt01	-	1642023	6
SCO1864	+	1998466	2	SCO1542	+	1651064	6
SCO2263	+	2432295	2	SCO1756	+	1877571	6
SCO5000	+	5438304	2	SCO1950	-	2085567	6
SCO5082	-	5523964	2	SCO2054	-	2202566	6
SCO7351	+	8163847	2	SCO2627	-	2854597	6
SCO7497	+	8308922	2	SCO3956	+	4355984	6
SCO0865	-	912865	3	SCO4201	-	4610678	6
SCO0867	-	914475	3	SCO5151	+	5598249	6
SCO1041	+	1097279	3	SCO5931	-	6500951	6
SCO1060	-	1119114	3	SCO7339	-	8154568	6
SCO1253	+	1324333	3	SCO0074	+	64987	7
SCO1601	+	1712155	3	SCO0881	-	928506	7
SCO1808	+	1939007	3	SCO1765	-	1885485	7
SCO2022	-	2168237	3	SCO1807	+	1936837	7
SCO2034	-	2186262	3	SCO1868	-	2002987	7
SCO2066	+	2216803	3	SCO1952	-	2087680	7
SCO2337	+	2508321	3	SCO2132	-	2293733	8
SCO2385	-	2557985	3	SCO2583	-	2789590	8
SCO2400	-	2572192	3	SCOt20	-	3079206	8
SCO2550	-	2751453	3	SCO3546	-	3920775	8
SCO2966	-	3227532	3	SCO6749	+	7505748	8

SCO0724	+	768417	9
SCO1684	-	1805151	9
SCO1860	-	1995066	9
SCO1903	+	2037021	9
SCO2657	+	2888288	9
SCO2870	+	3125325	9
SCO3978	-	4381958	9
SCO4264	-	4680245	9
SCO5048	+	5488235	9
SCO6499	+	7193821	9
SCO6618	-	7339958	9
SCO7015	-	7797133	9
SCO7093	+	7881832	9
SCO7729	+	8563616	9

Appendix II Transcriptional start sites in the secondary metabolite gene clusters of *S. coelicolor*.

TSS ID	Strand	TSS position	Category	Gene	secondary metabolite	Leaderless
TSS-0058	+	173768	PI	SCO0185	Isorenieratene	Leaderless
TSS-0059	-	182247	P	SCO0191	Isorenieratene	
TSS-0102	+	256148	P	SCO0268	Lantibiotic	
TSS-0103	+	256218	P	SCO0268	Lantibiotic	
TSS-0170	-	510997	PI	SCO0489	Coelichelin	
TSS-0171	+	511074	P	SCO0490	Coelichelin	
TSS-0172	-	511075	P	SCO0489	Coelichelin	
TSS-0173	-	527873	P	SCO0494	Coelichelin	
TSS-0174	-	532356	P	SCO0498	Coelichelin	
TSS-0175	+	532374	P	SCO0499	Coelichelin	
TSS-0176	+	532473	P	SCO0499	Coelichelin	
TSS-0266	+	796462	P	SCO0753	Bacteriocin	
TSS-0498	-	1339372	PI	SCO1268	PKS	Leaderless
TSS-0850	+	1998466	P	SCO1864	5-Hydroxyectoine	
TSS-0851	+	1998738	PI	SCO1865	5-Hydroxyectoine	
TSS-0852	+	2000421	P	SCO1866	5-Hydroxyectoine	
TSS-0853	+	2000764	PI	SCO1867	5-Hydroxyectoine	
TSS-1373	+	3035578	P	SCO2782	Desferrioxamine	
TSS-1654	-	3525930	P	SCO3215	CDA	
TSS-1655	+	3526080	P	SCO3216	CDA	
TSS-1656	+	3528805	P	SCO3217	CDA	
TSS-1657	+	3528878	P	SCO3217	CDA	
TSS-1658	+	3528912	P	SCO3217	CDA	
TSS-1659	-	3531481	P	SCO3218	CDA	
TSS-1660	+	3533028	A		CDA	
TSS-1661	+	3535978	A		CDA	
TSS-1662	-	3536808	PI	SCO3224	CDA	
TSS-1663	+	3538640	PI	SCO3226	CDA	
TSS-1664	-	3543133	P	SCO3229	CDA	
TSS-1665	+	3543259	P	SCO3230	CDA	
TSS-1666	-	3588722	P	SCO3236	CDA	
TSS-1667	-	3602370	P	SCO3248	CDA	
TSS-2727	-	5514249	PI	SCO5071	Actinorhodin	
TSS-2728	+	5514323	PI	SCO5072	Actinorhodin	Leaderless
TSS-2729	+	5516862	PI	SCO5075	Actinorhodin	
TSS-2730	+	5519054	PI	SCO5077	Actinorhodin	
TSS-2731	+	5519954	P	SCO5078	Actinorhodin	
TSS-2732	-	5523964	P	SCO5082	Actinorhodin	
TSS-2733	-	5524013	P	SCO5082	Actinorhodin	
TSS-2734	+	5524037	P	SCO5083	Actinorhodin	
TSS-2735	+	5528064	P	SCO5085	Actinorhodin	Leaderless

TSS-2736	-	5528518	A		Actinorhodin	
TSS-2737	-	5529751	P	SCO5086	Actinorhodin	
TSS-2738	+	5529894	PI	SCO5087	Actinorhodin	
TSS-2880	-	5786708	PI	SCO5315	Gray spore pigment	
TSS-2881	+	5790803	A		Gray spore pigment	
TSS-3172	+	6338652	P	SCO5799	Siderophore	
TSS-3220	+	6432601	PI	SCO5877	Prodiginine	
TSS-3221	+	6433591	PI	SCO5878	Prodiginine	
TSS-3222	-	6438972	P	SCO5881	Prodiginine	
TSS-3223	-	6440437	PI	SCO5882	Prodiginine	
TSS-3224	-	6444315	P	SCO5886	Prodiginine	
TSS-3225	+	6444567	P	SCO5888	Prodiginine	
TSS-3226	-	6462763	A		Prodiginine	
TSS-3227	-	6463084	A		Prodiginine	
TSS-3299	+	6666219	PI	SCO6073	Geosmin	Leaderless
TSS-3377	+	6891247	P	SCO6266	SCB1	
TSS-3382	-	6932046	P	SCO6275	Coelimycin P1	
TSS-3383	+	6932082	P	SCO6276	Coelimycin P1	
TSS-3384	+	6937977	P	SCO6280	Coelimycin P1	
TSS-3385	+	6939760	P	SCO6281	Coelimycin P1	
TSS-3386	-	6942568	P	SCO6282	Coelimycin P1	
TSS-3387	+	6942661	PI	SCO6283	Coelimycin P1	Leaderless
TSS-3388	+	6943643	P	SCO6284	Coelimycin P1	
TSS-3389	-	6946428	P	SCO6286	Coelimycin P1	
TSS-3390	+	6948543	P	SCO6289	Coelimycin P1	
TSS-3391	+	6951364	PI	SCO6292	Coelimycin P1	
TSS-3392	-	6954391	PI	SCO6294	Coelimycin P1	
TSS-3393	+	6954593	P	SCO6295	Coelimycin P1	
TSS-3454	+	7104846	P	SCO6429	Dipeptide	
TSS-3455	+	7116094	PI	SCO6433	Dipeptide	
TSS-3540	+	7422475	P	SCO6682	SapB	
TSS-3575	+	7515825	P	SCO6759	Hopene	
TSS-3576	-	7518821	A		Hopene	
TSS-3577	+	7520549	PI	SCO6764	Hopene	
TSS-3578	+	7523284	PI	SCO6766	Hopene	
TSS-3597	-	7598847	P	SCO6827	PKS	
TSS-3719	+	8026511	A		Germicidin	
TSS-3720	-	8027678	P	SCO7221	Germicidin	
TSS-3870	-	8506197	P	SCO7681	Coelibactin	

Appendix III RNA and RPF abundance of secondary metabolite genes in *S. coelicolor*.

Gene	Start	End	Strand	Length (bp)	RNA expression level				Ribosome occupancy				Product
					EE	ME	LE	S	RiboEE	RiboME	RiboLE	RiboS	
SCO0124	103628	104989	+	1362	4.56	6.14	65.15	65.28	0.91	0.00	0.00	0.84	hypothetical protein
SCO0125	105057	106640	+	1584	3.69	2.68	30.44	30.69	0.00	0.00	2.58	0.00	oxidoreductase
SCO0126	106637	112885	+	6249	22.41	29.14	187.74	181.11	1.82	2.80	1.72	10.98	beta keto-acyl synthase
SCO0127	112932	119654	+	6723	36.62	26.84	162.71	127.12	1.82	0.00	0.00	8.44	beta keto-acyl synthase
SCO0128	119651	119833	+	183	2.50	3.07	3.94	3.24	0.00	0.00	0.00	0.00	hypothetical protein
SCO0129	119841	120215	+	375	3.37	6.13	7.45	2.74	0.00	0.00	0.00	0.00	hypothetical protein
SCO0185	173768	174946	+	1179	782.88	188.95	43.17	215.99	56.52	29.40	9.47	58.26	geranylgeranyl pyrophosphate synthase
SCO0186	174943	176514	+	1572	762.63	178.98	71.05	266.44	32.82	4.20	13.77	41.38	phytoene dehydrogenase
SCO0187	176501	177496	+	996	304.28	88.53	33.45	153.51	19.14	11.20	5.16	20.27	phytoene synthase
SCO0188	177493	178503	+	1011	229.59	87.00	36.14	95.42	13.67	0.00	0.00	13.51	methyltransferase
SCO0189	178519	180087	-	1569	526.13	154.49	81.39	170.53	26.44	25.20	0.00	30.40	dehydrogenase
SCO0190	180084	180824	-	741	213.25	60.95	18.12	74.53	4.56	0.00	2.58	9.29	methyltransferase
SCO0191	180821	182038	-	1218	851.94	183.60	65.41	135.74	8.20	0.00	2.58	10.13	lycopen cyclase
SCO0267	255311	256087	+	777	12.38	11.50	15.56	17.57	0.91	0.00	0.00	0.00	hydrolase
SCO0268	256281	256442	+	162	20.08	334.01	396.02	932.51	0.91	228.18	194.50	221.23	hypothetical protein
SCO0269	256526	259687	+	3162	98.82	535.68	431.78	428.97	1.82	35.00	7.75	6.76	hypothetical protein
SCO0270	259684	261084	+	1401	21.66	78.60	76.33	64.44	0.00	0.00	0.86	0.00	hypothetical protein
SCO0489	510823	511035	-	213	42.25	97.76	18.64	288.35	14.59	19.60	1.72	43.91	hypothetical protein
SCO0490	511124	512215	+	1092	207.82	182.47	32.70	275.99	9.12	2.80	6.02	10.13	esterase
SCO0491	512302	513924	-	1623	329.03	368.41	183.18	792.83	18.23	12.60	3.44	15.20	ABC transporter
SCO0492	513989	524920	-	10932	1207.30	1321.09	356.53	5283.74	38.29	26.60	18.93	184.92	peptide synthetase
SCO0493	524978	526783	-	1806	1727.35	1600.90	424.90	2752.68	14.59	9.80	4.30	32.09	ABC-transporter transmembrane protein
SCO0494	526786	527838	-	1053	2255.71	2744.42	693.67	6598.67	203.28	134.39	46.47	362.25	iron-siderophore binding lipoprotein
SCO0495	527889	528752	-	864	699.41	710.73	322.77	1083.86	12.76	8.40	16.35	36.31	iron-siderophore ABC-transporter ATP-binding protein
SCO0496	528791	529897	-	1107	468.62	456.54	270.80	455.33	11.85	5.60	4.30	5.91	iron-siderophore permease transmembrane protein
SCO0497	529894	530970	-	1077	522.81	517.49	370.83	783.16	12.76	9.80	6.02	12.67	iron-siderophore permease transmembrane protein
SCO0498	530970	532325	-	1356	252.79	383.36	60.60	1921.01	28.26	5.60	16.35	90.35	peptide monooxygenase
SCO0499	532501	533448	+	948	246.79	291.72	50.25	949.95	76.57	58.79	24.10	76.84	formyltransferase
SCO0753	796584	796799	+	216	584.75	438.55	197.54	119.70	23.70	14.00	10.33	10.98	hypothetical protein
SCO0754	796977	797723	+	747	57.45	25.30	14.90	8.33	3.65	0.00	4.30	0.00	hypothetical protein
SCO0755	797720	799942	+	2223	94.94	48.68	21.42	15.28	3.65	0.00	0.00	0.84	ABC transporter
SCO0756	799944	802766	+	2823	142.70	70.92	47.80	18.08	0.00	0.00	0.86	0.84	ABC transporter
SCO1206	1277625	1278749	+	1125	97.95	92.40	123.39	157.19	1.82	11.20	6.02	2.53	polyketide synthase
SCO1207	1278746	1279960	+	1215	40.06	49.07	70.37	116.45	0.00	0.00	3.44	8.44	cytochrome P450
SCO1208	1279957	1280490	+	534	20.20	25.30	24.49	38.56	1.82	0.00	0.86	0.00	hypothetical protein
SCO1265	1335695	1336564	-	870	11.39	9.20	8.09	36.36	0.00	0.00	0.86	1.69	lipase
SCO1266	1336608	1337876	-	1269	2.70	5.37	4.16	31.96	0.00	0.00	0.00	0.84	3-oxoacyl-ACP synthase
SCO1267	1337873	1338130	-	258	2.38	1.92	2.98	11.80	0.00	0.00	0.00	0.00	acyl carrier protein
SCO1268	1338248	1339366	-	1119	5.12	8.43	6.61	27.53	0.00	2.80	0.00	0.00	acyltransferase
SCO1269	1339363	1340376	-	1014	4.76	8.43	9.15	22.21	0.00	0.00	0.00	0.84	pyruvate dehydrogenase subunit beta

SCO1270	1340373	1341344	-	972	5.00	5.75	9.37	25.69	0.00	0.00	2.58	1.69	pyruvate dehydrogenase subunit alpha
SCO1271	1341355	1342317	-	963	2.06	4.98	7.25	29.88	0.00	0.00	0.86	0.84	3-oxoacyl-ACP synthase
SCO1272	1342317	1342589	-	273	2.38	2.30	1.81	8.17	0.00	5.60	3.44	0.00	acyl carrier protein
SCO1273	1342661	1343779	-	1119	7.07	14.95	16.38	32.56	0.00	1.40	5.16	0.00	reductase
SCO1864	1998468	1998980	+	513	2448.42	1654.55	704.73	1339.59	84.78	103.59	37.87	26.18	acetyltransferase
SCO1865	1999110	2000381	+	1272	7998.87	5066.73	2826.70	5390.30	1270.75	988.30	429.45	454.29	diaminobutyrate--2-oxoglutarate aminotransferase
SCO1866	2000500	2000898	+	399	7966.01	6317.61	2233.20	2380.70	1000.01	722.32	260.77	156.21	ectC L-ectoine synthase
SCO1867	2000962	2001804	+	843	12271.87	10246.44	3719.90	3504.79	924.34	782.52	296.05	180.70	hydroxylase
SCO2700	2943456	2944322	-	867	7.38	14.19	253.59	125.46	0.00	0.00	0.86	5.91	tyrosinase
SCO2701	2944306	2944875	-	570	5.63	10.73	180.97	89.30	0.00	0.00	4.30	3.38	tyrosinase co-factor
SCO2782	3035641	3037083	+	1443	195.94	63.63	15.45	128.88	25.52	4.20	3.44	5.07	pyridoxal-dependent decarboxylase
SCO2783	3037106	3038347	+	1242	234.21	98.90	14.04	113.39	42.84	0.00	2.58	2.53	monooxygenase
SCO2784	3038344	3038898	+	555	85.01	34.12	2.76	36.01	16.41	1.40	4.30	0.84	acetyltransferase
SCO2785	3038895	3040682	+	1788	423.15	202.78	25.76	136.67	44.67	0.00	6.02	4.22	hypothetical protein
SCO3210	3519449	3520903	-	1455	29.84	42.17	828.57	1041.11	0.00	71.39	29.26	17.73	2-dehydro-3-deoxyheptonate aldolase
SCO3211	3520900	3521676	-	777	7.82	19.17	356.71	490.86	0.00	8.40	16.35	8.44	indoleglycerol phosphate synthase
SCO3212	3521673	3522680	-	1008	19.96	24.53	457.56	690.18	0.00	36.40	8.61	12.67	anthranilate phosphoribotransferase
SCO3213	3522697	3523299	-	603	3.81	8.05	242.48	324.18	0.00	7.00	0.86	3.38	anthranilate synthase component II
SCO3214	3523296	3524831	-	1536	26.71	34.50	644.97	834.03	3.65	0.00	12.05	11.82	anthranilate synthase component I
SCO3215	3524828	3525844	-	1017	26.10	34.89	762.58	988.43	0.00	57.39	52.50	39.69	hypothetical protein
SCO3216	3526137	3528527	+	2391	68.86	96.23	127.56	151.74	5.47	22.40	6.02	4.22	integral membrane ATPase
SCO3217	3529272	3531188	+	1917	231.35	771.02	3839.51	11131.65	35.55	618.74	112.74	357.18	transcriptional regulator
SCO3218	3531250	3531465	-	216	47.10	121.14	2408.44	2905.29	6.38	197.38	118.77	232.21	hypothetical protein
SCO3219	3531588	3532763	-	1176	29.16	69.77	155.92	244.88	3.65	0.00	0.00	3.38	lipase
SCO3220	3532971	3533399	-	429	35.91	36.80	546.97	1536.13	0.00	54.59	34.42	100.48	hypothetical protein
SCO3221	3533492	3534346	-	855	20.32	46.77	1412.06	2249.64	0.00	7.00	14.63	21.11	prephenate dehydrogenase
SCO3222	3534444	3534899	-	456	23.89	70.53	2088.83	4120.39	5.47	223.98	61.96	142.70	hypothetical protein
SCO3223	3535054	3535848	-	795	101.20	100.44	91.50	60.75	10.03	2.80	12.05	3.38	ABC transporter
SCO3224	3535855	3536808	-	954	130.80	128.43	135.35	76.55	4.56	21.00	9.47	6.76	ABC transporter ATP-binding protein
SCO3225	3536945	3538660	+	1716	415.20	412.49	283.73	213.29	11.85	12.60	6.02	3.38	two component sensor kinase
SCO3226	3538679	3539347	+	669	924.18	1105.61	861.17	476.67	59.25	76.99	46.47	26.18	two component system response regulator
SCO3227	3539337	3540671	-	1335	111.13	99.29	662.38	1239.07	0.91	48.99	24.10	32.93	aminotransferase
SCO3228	3540668	3541801	-	1134	20.15	22.62	524.74	990.87	0.00	0.00	5.16	13.51	glycolate oxidase
SCO3229	3541951	3543066	-	1116	14.20	44.86	1409.02	2721.10	1.82	123.19	61.10	107.24	4-hydroxyphenylpyruvic acid dioxygenase
SCO3230	3543335	3565726	+	22392	304.16	435.90	7031.48	8748.19	1.82	174.98	197.94	237.28	CDA peptide synthetase I
SCO3231	3565723	3576735	+	11013	230.69	260.31	5884.35	5260.38	14.59	18.20	224.62	179.86	CDA peptide synthetase II
SCO3232	3576735	3583988	+	7254	139.25	155.64	5063.12	3166.65	8.20	1.40	183.31	122.44	CDA peptide synthetase III
SCO3233	3583992	3584810	+	819	30.23	37.57	1363.96	1099.63	0.00	0.00	42.17	21.95	hydrolase
SCO3234	3584822	3585724	+	903	51.40	49.07	1674.10	1130.64	1.82	5.60	29.26	17.73	phosphotransferase
SCO3235	3585800	3587647	-	1848	38.53	62.10	2334.50	3090.72	0.00	28.00	35.29	38.84	ABC transporter
SCO3236	3587687	3588688	-	1002	17.51	57.12	2183.46	3315.52	0.00	37.80	85.20	89.51	oxygenase
SCO3237	3588746	3590134	-	1389	5.95	17.25	327.14	725.02	0.00	0.00	4.30	11.82	hypothetical protein
SCO3238	3590146	3591306	-	1161	4.13	11.88	284.66	612.08	2.73	0.00	5.16	13.51	hypothetical protein
SCO3239	3591313	3592182	-	870	8.45	14.57	361.09	820.25	3.65	5.60	6.02	15.20	hypothetical protein

SCO3240	3592179	3592889	-	711	4.76	11.88	267.99	386.27	0.00	0.00	4.30	3.38	hypothetical protein
SCO3241	3592886	3593758	-	873	7.02	13.80	442.33	522.62	0.00	14.00	18.07	17.73	hypothetical protein
SCO3242	3593755	3594630	-	876	10.08	16.10	444.50	664.47	0.00	8.40	2.58	2.53	putative transferase
SCO3243	3594627	3595793	-	1167	22.82	58.65	1643.45	2540.13	0.00	44.80	44.75	51.51	myo-inositol phosphate synthase
SCO3244	3595843	3596640	-	798	18.26	47.15	1376.41	3115.57	0.00	99.39	167.82	175.64	hypothetical protein
SCO3245	3596708	3597970	-	1263	22.82	32.20	649.36	1233.04	0.00	14.00	17.21	36.31	salicylate hydroxylase
SCO3246	3598017	3599009	-	993	4.69	23.38	559.26	1092.56	0.00	19.60	16.35	16.89	3-oxoacyl-ACP synthase
SCO3247	3599006	3600808	-	1803	17.63	57.12	1362.75	2959.12	0.00	58.79	20.65	76.84	acyl CoA oxidase
SCO3248	3600858	3602078	-	1221	29.14	74.38	2097.97	3291.80	0.00	12.60	25.82	33.78	3-oxoacyl-ACP synthase
SCO3249	3602075	3602320	-	246	13.82	40.26	897.20	1484.12	0.91	260.37	30.12	101.33	acyl carrier protein
SCO5071	5513809	5514249	-	441	0.87	2.30	28.32	105.55	0.00	0.00	0.00	10.13	hydroxylacyl-CoA dehydrogenase
SCO5072	5514323	5515246	+	924	7.82	19.55	180.69	380.73	0.00	0.00	0.86	18.58	hydroxylacyl-CoA dehydrogenase
SCO5073	5515243	5516232	+	990	18.84	17.64	172.44	368.94	0.00	1.40	2.58	12.67	oxidoreductase
SCO5074	5516299	5516943	+	645	13.77	18.02	341.82	796.34	1.82	1.40	30.12	46.44	dehydratase
SCO5075	5516977	5517897	+	921	62.50	52.91	353.50	591.46	0.91	0.00	1.72	10.98	oxidoreductase
SCO5076	5517894	5519495	+	1602	293.89	340.44	539.96	906.40	4.56	0.00	8.61	13.51	hypothetical protein
SCO5077	5519497	5519892	+	396	160.08	218.91	450.33	1100.24	20.05	36.40	45.61	84.44	hypothetical protein
SCO5078	5519987	5520832	+	846	179.46	203.57	465.17	939.02	4.56	1.40	1.72	17.73	hypothetical protein
SCO5079	5520857	5521741	+	885	51.09	66.33	297.81	618.50	7.29	8.40	7.75	17.73	hypothetical protein
SCO5080	5521738	5522883	+	1146	59.12	52.90	263.69	546.94	0.91	0.00	14.63	22.80	hydrolase
SCO5081	5522876	5523217	+	342	44.80	46.77	126.21	232.83	0.00	2.80	0.00	3.38	hypothetical protein
SCO5082	5523183	5523962	-	780	792.53	739.49	630.55	687.48	27.35	37.80	17.21	17.73	transcriptional regulator
SCO5083	5524073	5525809	+	1737	284.33	208.91	138.36	616.35	10.03	0.00	1.72	33.78	actinorhodin transporter
SCO5084	5525809	5527944	+	2136	249.75	211.61	126.92	528.51	13.67	5.60	6.88	15.20	hypothetical protein
SCO5085	5528094	5528861	+	768	131.26	137.25	247.35	1034.22	0.00	11.20	7.75	40.53	actinorhodin cluster activator protein
SCO5086	5528935	5529720	-	786	13.77	10.35	33.78	417.66	0.00	0.00	0.00	26.18	ketoacyl reductase
SCO5087	5529801	5531204	+	1404	15.52	18.40	116.84	392.46	0.00	0.00	2.58	17.73	actinorhodin polyketide beta-ketoacyl synthase subunit alpha
SCO5088	5531201	5532424	+	1224	13.33	14.57	86.94	265.19	0.00	0.00	0.00	9.29	actinorhodin polyketide beta-ketoacyl synthase subunit beta
SCO5089	5532449	5532709	+	261	4.01	5.37	17.04	93.89	6.38	0.00	0.86	5.91	actinorhodin polyketide synthase ACP
SCO5090	5532706	5533656	+	951	20.98	20.70	62.25	206.78	0.00	0.00	5.16	5.07	actinorhodin polyketide synthase bifunctional cyclase/dehydratase
SCO5091	5533653	5534546	+	894	27.58	22.23	76.05	148.42	0.91	0.00	0.00	5.91	cyclase
SCO5092	5534558	5535091	+	534	18.28	14.57	39.95	93.90	0.00	0.00	0.00	3.38	actinorhodin polyketide dimerase
SCO5222	5681016	5682101	+	1086	155.42	370.79	484.72	311.11	0.00	180.58	65.41	15.20	lyase
SCO5223	5682098	5683483	+	1386	145.56	340.86	579.16	415.66	11.85	128.79	105.86	14.35	cytochrome P450
SCO5314	5785753	5786088	-	336	23.41	22.62	28.23	45.29	0.91	0.00	0.86	0.00	whiE protein VII
SCO5315	5786122	5786601	-	480	20.52	26.84	39.64	74.55	0.91	0.00	0.00	3.38	polyketide cyclase
SCO5316	5786603	5786875	-	273	10.63	9.58	17.26	30.50	0.00	4.20	5.16	3.38	acyl carrier protein
SCO5317	5786945	5788219	-	1275	12.14	14.57	26.22	30.77	1.82	0.00	0.86	0.00	polyketide beta-ketoacyl synthase beta
SCO5318	5788216	5789487	-	1272	9.76	12.27	23.52	28.07	0.91	0.00	0.00	1.69	polyketide beta-ketoacyl synthase alpha
SCO5319	5789484	5789957	-	474	4.44	1.53	8.31	9.63	0.00	0.00	1.72	0.00	whiE protein II
SCO5320	5790104	5791297	-	1194	19.16	19.55	24.90	34.55	1.82	0.00	0.00	2.53	whiE protein I
SCO5799	6338949	6340547	+	1599	32.17	29.13	398.80	324.37	0.00	0.00	12.91	11.82	aminotransferase
SCO5800	6340587	6342533	+	1947	7.26	16.49	295.26	271.04	0.91	0.00	6.88	13.51	hypothetical protein
SCO5801	6342563	6343258	+	696	11.27	9.97	160.15	145.25	0.00	0.00	3.44	7.60	hypothetical protein

SCO5877	6432566	6433618	+	1053	148.79	1049.85	5382.42	9739.06	5.47	181.98	64.55	269.36	transcriptional regulator RedD
SCO5878	6433668	6436616	+	2949	225.08	1186.31	6973.17	12660.45	2.73	149.78	135.12	324.25	polyketide synthase RedX
SCO5879	6436613	6437788	+	1176	125.80	729.28	4136.54	7102.07	12.76	123.19	97.25	189.99	acyl-coa dehydrogenase RedW
SCO5880	6437814	6438134	+	321	50.14	278.75	1682.25	3334.79	2.73	32.20	75.73	200.12	RedY protein
SCO5881	6438206	6438859	-	654	695.60	1524.04	1599.27	4463.13	10.94	128.79	27.54	107.24	response regulator
SCO5882	6439123	6440310	-	1188	8.89	107.36	780.56	1032.68	0.91	2.80	17.21	54.89	RedV protein
SCO5883	6440423	6441208	-	786	6.63	134.97	976.42	1203.87	0.00	0.00	4.30	20.27	hypothetical protein
SCO5884	6441276	6442166	-	891	17.77	289.12	1864.10	2830.30	0.00	55.99	35.29	105.55	hypothetical protein
SCO5885	6442174	6442614	-	441	8.57	188.28	969.30	1589.77	0.00	37.80	24.96	73.46	hypothetical protein
SCO5886	6442753	6443976	-	1224	37.88	942.51	6926.33	8674.49	0.00	85.39	54.22	189.99	3-oxoacyl-ACP synthase
SCO5887	6443973	6444218	-	246	14.33	478.94	2970.56	4301.31	3.65	1213.67	142.86	1520.77	acyl carrier protein
SCO5888	6444587	6445594	+	1008	31.42	776.10	4266.78	5366.02	0.00	116.19	35.29	172.26	3-oxoacyl-ACP synthase
SCO5889	6445654	6445917	+	264	10.83	270.33	1387.12	2670.12	0.00	57.39	37.01	80.22	hypothetical protein
SCO5890	6445914	6447836	+	1923	52.16	1012.31	6205.45	10379.87	0.91	172.18	123.93	362.25	8-amino-7-oxononanoate synthase
SCO5891	6447950	6449548	+	1599	22.65	437.51	2899.38	4965.79	1.82	170.78	76.59	241.50	peptide synthase
SCO5892	6449549	6456442	+	6894	47.81	973.13	8409.36	13668.88	4.56	254.77	236.67	480.47	polyketide synthase
SCO5893	6456446	6457489	+	1044	14.33	236.96	2913.47	4374.89	0.00	43.40	81.76	163.81	oxidoreductase
SCO5894	6457476	6458318	+	843	15.20	195.55	2180.49	3152.03	0.91	21.00	34.42	107.24	thioesterase
SCO5895	6458306	6459394	+	1089	23.96	378.82	4798.97	7543.68	0.91	106.39	166.10	327.63	methyltransferase
SCO5896	6459417	6462218	+	2802	60.82	799.45	11322.77	16323.28	7.29	208.58	321.01	586.86	phosphoenolpyruvate-utilizing enzyme
SCO5897	6462296	6463483	+	1188	111.84	703.98	11862.90	15259.84	3.65	194.58	427.73	696.63	oxidase
SCO5898	6463526	6464206	+	681	114.24	519.92	8232.46	7356.84	8.20	135.79	324.45	276.12	hypothetical protein
SCO6073	6666219	6668399	+	2181	171.81	516.44	4775.96	1094.99	29.17	71.39	308.10	81.91	cyclase
SCO6266	6891293	6892237	+	945	66.33	195.53	2007.09	567.39	10.03	51.79	163.52	44.75	ScbA protein
SCO6273	6900898	6907356	-	6459	209.54	178.28	106483.67	9328.68	22.79	5.60	7219.71	390.11	type I polyketide synthase
SCO6274	6907413	6918143	-	10731	120.58	143.79	81246.25	8453.81	13.67	5.60	3933.88	338.61	type I polyketide synthase
SCO6275	6918286	6931959	-	13674	117.32	265.37	98939.70	9258.92	11.85	4.20	2999.25	519.31	type I polyketide synthase
SCO6276	6932285	6933607	+	1323	157.12	302.93	126410.91	25687.01	36.46	18.20	5433.93	1330.78	hypothetical protein
SCO6277	6933604	6934464	+	861	121.86	213.95	77380.54	14563.55	5.47	2.80	1083.52	260.92	epoxide hydrolase
SCO6278	6934507	6936138	+	1632	276.65	434.82	177702.38	31048.30	40.11	19.60	9930.66	1254.78	integral membrane transport protein
SCO6279	6936149	6937705	+	1557	207.86	362.35	161827.57	25893.14	41.02	7.00	6784.24	945.73	diaminobutyrate-pyruvate aminotransferase
SCO6280	6938012	6939643	+	1632	143.50	287.15	14388.20	11545.34	5.47	12.60	357.16	595.30	regulatory protein
SCO6281	6939907	6941544	+	1638	39.97	96.24	12405.38	3907.31	2.73	1.40	271.09	192.52	FAD-binding protein
SCO6282	6941746	6942543	-	798	922.22	1426.72	387172.27	178501.90	146.76	62.99	25227.26	9355.99	3-oxoacyl-ACP reductase
SCO6283	6942661	6943506	+	846	216.12	333.58	61598.73	15838.97	38.29	4.20	2420.91	882.40	hypothetical protein
SCO6284	6943673	6945265	+	1593	147.07	241.94	45065.34	8509.72	26.44	22.40	2450.18	186.61	decarboxylase
SCO6285	6945280	6945543	+	264	19.52	30.29	6197.88	1235.16	0.00	12.60	434.61	60.80	hypothetical protein
SCO6286	6945705	6946379	-	675	66.26	75.53	3061.44	4160.06	0.00	2.80	561.98	464.42	regulatory protein
SCO6287	6946617	6947423	+	807	22.53	28.37	4921.61	323.12	0.91	0.00	131.67	16.04	thioesterase II
SCO6288	6947584	6948414	+	831	22.65	34.89	6574.01	867.36	0.00	0.00	98.11	35.46	regulatory protein
SCO6429	7104902	7106290	+	1389	22.46	115.40	114.01	724.91	0.00	4.20	0.00	21.11	hypothetical protein
SCO6430	7106284	7108176	+	1893	30.64	166.39	80.44	502.74	0.00	0.00	1.72	6.76	hypothetical protein
SCO6431	7108264	7111779	+	3516	44.72	292.11	142.43	1001.04	10.94	15.40	27.54	82.75	peptide synthase
SCO6432	7111866	7116089	+	4224	47.81	200.87	95.20	446.60	0.00	0.00	4.30	24.49	peptide synthase

SCO6433	7116086	7116493	+	408	11.39	62.11	33.65	300.73	0.91	4.20	3.44	16.04	hypothetical protein
SCO6434	7116513	7117874	+	1362	31.15	169.45	103.89	731.26	0.00	0.00	4.30	39.69	oxidoreductase
SCO6435	7117871	7118257	+	387	13.45	44.47	38.32	237.40	2.73	0.00	0.00	14.35	hypothetical protein
SCO6436	7118254	7119774	+	1521	26.39	127.27	65.81	495.93	0.91	4.20	0.00	25.33	tRNA synthetase
SCO6437	7119848	7121122	+	1275	21.08	110.40	59.74	405.02	8.20	11.20	10.33	26.18	hypothetical protein
SCO6438	7121125	7122447	+	1323	354.13	502.57	274.95	808.06	25.52	15.40	37.87	41.38	diaminopimelate decarboxylase
SCO6681	7419664	7422456	+	2793	11.24	32.20	430.96	1102.01	1.82	7.00	22.38	34.62	Ser/Thr protein kinase
SCO6682	7422494	7422622	+	129	32.90	109.23	1475.32	12863.70	15.50	86.79	1334.82	752.36	hypothetical protein
SCO6683	7422658	7424568	+	1911	10.63	31.81	268.82	1145.43	0.00	0.00	6.88	10.98	ABC transporter ATP-binding protein
SCO6684	7424565	7426391	+	1827	7.14	6.13	53.05	115.42	0.00	0.00	1.72	0.00	ABC transporter ATP-binding protein
SCO6685	7426396	7427004	-	609	9.01	14.18	89.92	228.74	0.00	2.80	1.72	13.51	two-component system response regulator
SCO6759	7516017	7516928	+	912	709.65	485.68	680.11	813.75	16.41	19.60	24.96	29.55	phytoene synthase
SCO6760	7516937	7517875	+	939	720.84	542.01	730.99	802.81	23.70	0.00	21.52	12.67	phytoene synthase
SCO6761	7517875	7518033	+	159	265.34	177.86	230.54	223.18	1.82	0.00	1.72	4.22	hypothetical protein
SCO6762	7518030	7519466	+	1437	1399.03	855.58	1094.02	1226.08	17.32	7.00	41.31	21.95	phytoene dehydrogenase
SCO6763	7519463	7520599	+	1137	1621.38	1100.88	1087.98	1134.00	76.57	40.60	57.66	32.09	polyprenyl synthetase
SCO6764	7520716	7522758	+	2043	760.97	882.50	795.93	1848.42	18.23	25.20	71.43	56.58	squalene-hopene cyclase
SCO6765	7522758	7523399	+	642	82.95	125.36	139.50	311.98	4.56	1.40	13.77	9.29	lipoprotein
SCO6766	7523406	7524428	+	1023	214.45	458.93	769.42	1555.57	33.73	18.20	252.16	144.39	hypothetical protein
SCO6767	7524433	7525587	+	1155	133.98	241.92	397.47	849.71	3.65	0.00	4.30	9.29	ispG 4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase
SCO6768	7525613	7527583	+	1971	140.56	325.90	568.96	1663.95	0.91	8.40	50.78	64.17	1-deoxy-D-xylulose-5-phosphate synthase
SCO6769	7527584	7528969	+	1386	107.69	256.12	497.68	1290.41	24.61	25.20	43.89	68.40	aminotransferase
SCO6770	7529043	7529657	+	615	879.95	525.92	333.37	485.72	6.38	22.40	23.24	30.40	DNA-binding protein
SCO6771	7529654	7529773	+	120	186.98	179.78	110.39	224.76	1.82	35.00	30.98	41.38	small hydrophobic hypothetical protein
SCO6826	7590412	7591452	-	1041	10.95	11.12	13.53	31.84	1.82	0.00	3.44	0.00	hypothetical protein
SCO6827	7591479	7598555	-	7077	45.79	51.76	58.77	121.40	9.12	5.60	25.82	7.60	polyketide synthase
SCO6927	7691781	7692905	-	1125	66.17	56.36	82.55	113.01	3.65	0.00	0.00	3.38	hypothetical protein
SCO6928	7692907	7694997	-	2091	153.70	186.69	312.65	461.54	24.61	8.40	24.10	11.82	O-methyltransferase
SCO6929	7695016	7696275	-	1260	8.77	12.65	20.56	46.02	0.91	1.40	0.00	0.84	hypothetical protein
SCO6930	7696262	7699360	-	3099	53.17	55.21	85.38	122.91	2.73	0.00	6.88	2.53	hypothetical protein
SCO6931	7699447	7699626	-	180	8.77	10.73	65.74	158.02	0.00	4.20	1.72	4.22	hypothetical protein
SCO6932	7699664	7699795	-	132	9.71	9.58	44.03	97.42	0.00	0.00	0.00	5.91	hypothetical protein
SCO7221	8026306	8027475	-	1170	56.79	20.32	133.74	548.22	8.20	0.00	23.24	59.11	polyketide synthase
SCO7669	8493549	8494580	-	1032	1.94	1.53	0.00	0.95	1.82	0.00	0.00	2.53	oxidoreductase
SCO7670	8494577	8495092	-	516	0.44	0.38	0.53	0.95	0.00	0.00	0.00	0.00	hypothetical protein
SCO7671	8495102	8496340	-	1239	0.75	0.77	2.23	0.00	0.00	0.00	0.00	0.00	transferase
SCO7681	8504461	8506122	-	1662	23.89	23.39	24.18	22.66	0.91	5.60	0.00	0.00	AMP-binding ligase
SCO7682	8506283	8512972	+	6690	70.71	52.52	25.96	35.49	10.03	1.40	0.86	3.38	non-ribosomal peptide synthase
SCO7683	8512969	8518497	+	5529	60.63	49.83	34.60	59.82	1.82	0.00	0.86	0.00	non-ribosomal peptide synthase
SCO7684	8518494	8519645	+	1152	8.77	9.20	7.87	6.54	0.00	0.00	0.00	0.00	hypothetical protein
SCO7685	8519642	8521675	+	2034	11.51	7.67	8.09	8.35	0.00	0.00	0.00	3.38	hypothetical protein
SCO7686	8521684	8522919	+	1236	11.58	10.73	14.06	19.38	0.91	21.00	0.86	0.00	cytochrome P450
SCO7687	8522916	8523749	+	834	10.51	4.60	9.37	9.63	0.00	0.00	0.00	0.84	thioesterase
SCO7688	8523760	8524488	+	729	8.18	8.05	10.01	9.07	0.00	0.00	0.00	0.84	hypothetical protein

SCO7689	8524521	8526347	+	1827	17.89	21.85	22.79	17.84	0.00	0.00	0.00	0.00	ABC transporter ATP-binding protein
SCO7690	8526344	8528107	+	1764	17.26	21.85	16.09	19.18	0.00	0.00	2.58	2.53	ABC transporter ATP-binding protein
SCO7691	8528166	8529581	+	1416	28.09	18.40	27.28	30.11	0.91	1.40	2.58	3.38	lyase
SCO7700	8537030	8538352	+	1323	9.08	13.80	452.56	779.59	0.00	0.00	8.61	15.20	cyclase
SCO7701	8538368	8539246	+	879	20.91	24.54	748.26	1038.47	0.91	0.00	18.93	73.46	methyltransferase

국문 초록

본 연구는 차세대 시퀀싱 기술의 여러 가지 응용방법을 이용한 방선균 유전체의 구조적, 기능적 특성 및 세포 내 발현변화의 분석을 목표로 하고 있다.

첫째로, 유전체 시퀀싱이 완료된 17 종의 방선균 비교유전체 분석으로 모든 종에 공통으로 존재하는 핵심 유전자를 선별할 수 있었다. 이를 통해 2018개의 상동유전자 집단이 핵심 유전체임을 밝히고 그 중에서 스트레스 방어에 관련된 이차대사 생산 유전자들이 존재함을 확인하였다.

다음으로, ChIP-seq을 통한 *Streptomyces coelicolor*의 전사조절네트워크 규명 기술을 확립하였다. 이를 위해 다양한 단백질에 응용 가능한 연쇄 myc 항원의 유전체 삽입 방식을 개발하여 NdgR 조절네트워크 분석에 이용하였다. 크로마틴 면역침강법으로 확보한 NdgR의 타겟 DNA를 차세대 시퀀싱 방법으로 분석하여 총 19개의 결합위치와 보존 서열을 밝혔다. 이를 통해 NdgR이 분지 아미노산 및 황 결합 아미노산의 합성유전자를 조절함을 알았으며 특히 피드포워드 제어를 통해 산화 스트레스 조건에서 황 동화 작용의 항상성을 유지하는 역할을 하는 것을 밝혔다.

마지막으로 *S. coelicolor* 전사체의 구조와 세포의 성장에 따른 발현변화를 유전체 수준에서 확인할 수 있었다. TSS-seq으로 총 3926개의 전사시작위치를 밝히고 유전자별 비번역 부위의 길이를 파악할 수 있

었다. 이를 통해 약 20%의 유전자가 leaderless 유전자임을 알아냈고 그 중 가장 많은 비율이 전사 관련 기능을 하는 것을 확인하였다. 특히 21개의 이차대사 클러스터에 존재하는 80개의 전사시작 위치를 규명하여 전사단위의 정확한 유전자 조작에 필요한 정보를 제공하였다. 또한 Ribo-seq 데이터와의 비교를 통해 번역 수준에서의 조절에 의한 RNA 발현과 단백질 발현의 불일치를 확인하였고 이차대사 유전자들 각각의 전사 및 번역을 정량하여 유전자별 번역효율을 파악할 수 있었다.

본 연구를 통해 확립된 방선균 유전체의 다양한 정보는 항생제를 비롯한 산업적으로 중요한 이차대사물 생산의 시스템 수준에서의 이해와 유전자 조작에 필요한 중요한 자료가 될 것으로 기대한다.

주요어: 방선균, 비교유전체학, 전사체학, 전사조절, 차세대 시퀀싱

학번: 2007-21181